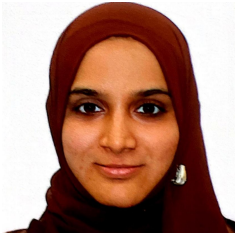
Team A Project
Case Study
Phase 1

INTRODUCTION



Objective of Case Study

The Case Study is intended to show the team's journey using a bank setting to address a business issue. The process entails a walkthrough - a structured process of understanding the business issue, data understanding to data preparation process in a logical manner to support the ultimate findings and proposing a logical, realistic solution addressing the objectives defined.

Participants

The Discussion panel was attended by the following stakeholders:

SAP Project Team

Project Manager

Data Scientist

Technology Consultant

Customers

Project Sponsor

Sale/Marketing Manager

Regional HR Director

Data Science Expert

Project Methodology- CRISP-DM

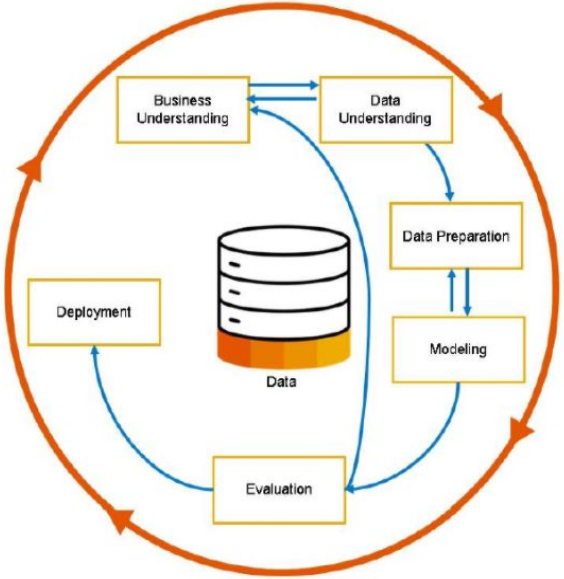
Cross-Industry Standard Process for Data Mining (CRISP-DM)

Phase1:

- Business Understanding
- Data Understanding
- Data Preparation

Phase 2:

- Modeling
- Evaluation
- Deployment



extract: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

CRISP-DM Project Methodology

35 days

TIME						
Project Prep & readiness	5 days					
Business Understanding	6 days					
Data Understanding		11 days				
Modeling			6 days			
Evaluation					2 days	
Deployment						5 days

Project Plan (18 Feb- 14 Apr'22)

		Name	Duration	Start	Finish	Predecessors
2		**Industry Project**	**40 days?**	**18/2/22 8:00 AM**	**14/4/22 5:00 PM**	
3		**Project Preparation and Readiness**	**5 days?**	**18/2/22 8:00 AM**	**24/2/22 5:00 PM**	
4		Identify Project Team and Stakeholders	1 day?	18/2/22 8:00 AM	18/2/22 5:00 PM	
5		Meeting Invites	1 day?	21/2/22 8:00 AM	21/2/22 5:00 PM	
6		Kickoff meeting & Acquire Executive sponsorship	1 day?	22/2/22 8:00 AM	22/2/22 5:00 PM	5
7		Check system infrastucture and applications	2 days?	23/2/22 8:00 AM	24/2/22 5:00 PM	6
8		**Business Understanding**	**11 days?**	**18/2/22 8:00 AM**	**4/3/22 5:00 PM**	
9		Determine Business Objectives	1 day?	18/2/22 8:00 AM	18/2/22 5:00 PM	
10		Assess situation	1 day?	18/2/22 8:00 AM	18/2/22 5:00 PM	
11		Determine Data Science Goals & criteria	1 day?	18/2/22 8:00 AM	18/2/22 5:00 PM	
12		Brainstorming	1 day?	21/2/22 8:00 AM	21/2/22 5:00 PM	11
13		Project Plan	9 days?	22/2/22 8:00 AM	4/3/22 5:00 PM	12
14		**Data Understanding**	**10 days?**	**3/3/22 8:00 AM**	**16/3/22 5:00 PM**	
15		Collect initial Data	6 days?	3/3/22 8:00 AM	10/3/22 5:00 PM	
16		Describe Data	1 day?	11/3/22 8:00 AM	11/3/22 5:00 PM	
17		Explore Data	1 day?	14/3/22 8:00 AM	14/3/22 5:00 PM	15
18		Verify Data Quality	3 days?	14/3/22 8:00 AM	16/3/22 5:00 PM	15
19		**Data Preparation**	**10 days?**	**16/3/22 8:00 AM**	**29/3/22 5:00 PM**	
20		Select Data	2 days?	16/3/22 8:00 AM	17/3/22 5:00 PM	
21		Clean Data	2 days?	18/3/22 8:00 AM	21/3/22 5:00 PM	20
22		Construct Data	3 days?	18/3/22 8:00 AM	22/3/22 5:00 PM	20
23		Integrate Data	2 days?	23/3/22 8:00 AM	24/3/22 5:00 PM	22
24		Format Data	3 days?	25/3/22 8:00 AM	29/3/22 5:00 PM	23
25		**Modeling**	**6 days?**	**30/3/22 8:00 AM**	**6/4/22 5:00 PM**	**19**
26		Select Modeling Technique	1 day?	30/3/22 8:00 AM	30/3/22 5:00 PM	
27		Generate Test Design	1 day?	31/3/22 8:00 AM	31/3/22 5:00 PM	26
28		Build Model	1 day?	31/3/22 8:00 AM	31/3/22 5:00 PM	26
29		Assess Model	4 days?	1/4/22 8:00 AM	6/4/22 5:00 PM	28
30		**Evaluation**	**2 days?**	**7/4/22 8:00 AM**	**8/4/22 5:00 PM**	**25**
31		Evaluate Results	1 day?	7/4/22 8:00 AM	7/4/22 5:00 PM	
32		Review	1 day?	8/4/22 8:00 AM	8/4/22 5:00 PM	31
33		Determine Next Steps	1 day?	8/4/22 8:00 AM	8/4/22 5:00 PM	31
34		**Deployment**	**5 days?**	**8/4/22 8:00 AM**	**14/4/22 5:00 PM**	
35		Plan Deployment	1 day?	8/4/22 8:00 AM	8/4/22 5:00 PM	
36		Plan Monitoring & Maintenance	1 day?	8/4/22 8:00 AM	8/4/22 5:00 PM	
37		Produce Final Report	2 days?	11/4/22 8:00 AM	12/4/22 5:00 PM	35

Bank's IT Maturity Assessment

Maturity Model	Adhoc	Prepared	Demonstrated	Proven	Intelligent
Strategy	Business units do not consider data driven policy & practice	Business units have vision for data driven policy & practice	Business units have strategy for data driven policy and practice documented and endorsed	Business units have strategy formulated based on data-driven insights	Org strategy is formulated based on data-driven insights
Data	Data is unavailable to support data driven transformation	Data sources have been identified and data is being prepared	Quality data is available in some domains	Data is managed as strategic asset across org	Data is continually enriched by data-driven processes
Technology	Soloed systems impede longitudinal analysis	Longitudinal analysis is supported by traditional analytics	Point solutions supported predictive analtics in some dimensions	System Platform supports predictive analytics across domains	Org includes automatic refinement of predictive models
Culture	Culture resistance to change	Executive support for data driven policy	Partial adoption of data driven practice	Org adopts data practice widely	Data driven culture extends to Org decision matters
Influence	Decisions are based on gut instinct	Decisions are based on unreliable and incomplete data	Some evidence of data decision making	Business practices are transformed based on data driven insights	Org practices are influeced by data
Impact	Business practices are unchange	Policy & practice reviews being promoted by data-driven insights	Some examples of dynamic responses to major events and emerging trends	Dynamic responses are frequently informed by data-driven insights	Org practices are continually reviewed and improved

Business Understanding

Business Understanding Tasks

Determine business objectives

Assess Situation

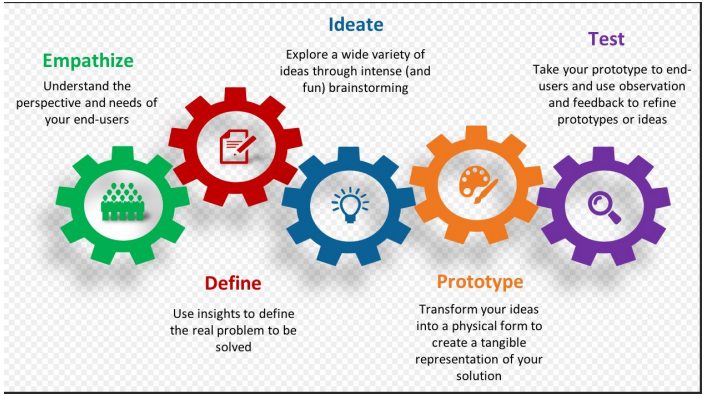
Determine analytics goals

Produce a detailed project plan

DESIGN THINKING METHODOLOGY

Understanding the key needs of the bank

- Increase revenue through higher credit card expenditure and interest income through property loans
- Improve effectiveness of campaigns and reduce labour costs
- Reduce bad debt due to credit card defaults with early intervention



POINT-OF-VIEW

Requirements:
Understand business needs and select most important ones:

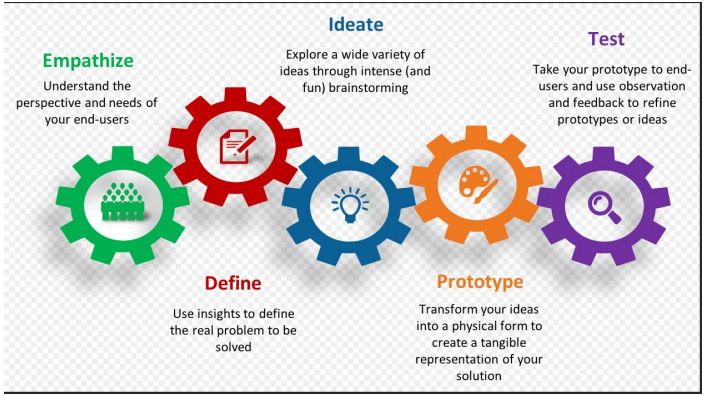
- Launch new Credit Cards
- Reduce credit risk on re-payments
- Launch effective communications strategy
- Promote other product Property Loans to potential good customers
- Improve Deposits position.

Insights:
Express the Insights:

- The Bank already has a loyalty scheme for its retail account holders / customers
- The customers are spending using their Savings Bank accounts
- Customers are looking for Property Loans
- Bank is seeing a decreasing trend in Deposits.

Point-of-view:
A point-of-view combines needs and insights:

- Existing customers are using their Savings bank account for retail purchases
- Potential to introduce Credit Cards that offer better offers and facilities
- Cross-sell other products
- Attract customers to deposit schemes.



Data Understanding

Data Understanding Tasks

Collect Initial Data

↓

Explore and describe the data

↓

Verify data quality and identify outlier

Introduction to HANA

What is HANA?

- Modeling refers to an activity of refining or slicing data in database tables by creating information views that depict a business scenario. These information views for any reporting and decision making purposes.

Who uses?

- All organisations which has procurements, sales, QA QC, etc. Cloud providers i.e. AWS, MS Azure, Google Cloud Platform, etc.

Where is HANA applied?

- In every parts within the organisation.

When is HANA used ?

- Companies that process massive amounts of data with near zero latency, query data in an instant, and become truly data driven.

How does HANA works?

- It is designed to replicate and ingest structured data from SAP and non-SAP relational databases, applications, and other systems quickly.

Why? What are the reasons for using HANA?

- SAP HANA Modeler option is used to create Information views on the top of schemas → tables in HANA database.

Advantages: One copy of data is needed.

Disadvantages: For cloud users, downtime during SAP maintenance.

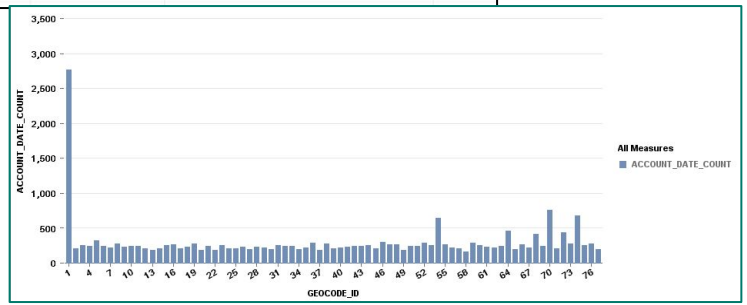
LIST OF TABLES

1. CS_ACCOUNT
2. CS_BRANCH
3. CS_BRANCHMGR
4. CS_CARD_TRANSACTIONS
5. CS_CITY
6. CS_CLIENT
7. CS_CLIENT_CONTACT
8. CS_CONTACTTYPE
9. CS_CREDIT_CARD
10. CS_DEPT
11. CS_DEPTTEXTMODE
12. CS_DESIGNATION
13. CS_DESIGNATIONTEXT
14. CS_EMPLOYEE
15. CS_EMP_CTC
16. EMP_SALCHANGE
17. CS_GEODEMO
18. CS_GLENTRY
19. CS_GL_CONFIG
20. CS_LOAN
21. CS_LOYALTY_CARD
22. CS_MAPPINGTABLE
23. CS_MODE
24. CS_PRODUCT
25. CS_PRODUCTGROUP
26. CS_TRANSACTIONS
27. CS_TRANSACTIONS_BACKEND

- Explore using HANA Studio
- Check data quality
- Identify fields and their data types
- Find out number of records

CS Account

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	ACCOUNT_ID	INTEGER		INT	X(1)	X		Account Number
2	GEOCODE_ID	INTEGER		INT				Geocode Number
3	FREQUENCY	NVARCHAR	9	STRING				Frequency
4	ACCOUNT_DATE	DATE		DAYDATE				Account Opening Date



Total Records:22,500

Primary Key:ACCOUNT_ID

Foreign Key:GEOCODE_ID

No missing values

No Negative values

Account date from from 01 Jan 2012 to 22 Oct 2016

Most of the customers in GEOCODE_ID-1-2,770 Records

Most of the customers perform monthly transactions.

Highest 'FREQUENCY' is Monthly-20,835 Records

AB FREQUENCY	12 ACCOUNT_DATE_COUNT
Immediate	465
Monthly	20,835
Weekly	1,200

CS Branch

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	BRANCHID	INTEGER		INT	X(1)	X		Branch Identifier
2	NAME	NVARCHAR	40	STRING				Branch Name
3	CITYID	INTEGER		INT				City ID
4	STARTYEAR	INTEGER		INT				Branch Start Date
5	MAINBRANCH	NVARCHAR	1	STRING				City/Region Main Branch
6	BRANCHTYPE	NVARCHAR	4	STRING				Level/Type of Branch
7	REPBRANCH	INTEGER		INT				Reporting / Managing Branch

Total Number of Records:48

Primary Key:BRANCHID

Foreign Key:CITYID

One missing value for New York A BRANCHID 7

No Negative values

Start year from 1954-2009

Head Office at New York A

BRANCHID	NAME	CITYID	STARTYEAR	MAINBRANCH	BRANCHTYPE	REPBRANCH
7	New York A	16	1,999	Y	HO	?
26	Boston A	2	1,992	Y	BR	7
21	Boston B	2	1,998	N	BR	7
40	Charlotte A	3	2,001	Y	BR	7
39	Charlotte B	3	1,982	N	BR	7
36	Chicago	4	1,954	N	RE	7
42	New York B	16	1,985	N	BR	7
10	New York C	16	1,978	N	BR	7
31	Philadelphia A	18	1,984	Y	BR	7
2	Philadelphia B	18	1,967	N	BR	7
9	San Francisc...	22	1,957	Y	RE	7
29	Washington	25	1,996	N	BR	7
30	Dallas A	6	1,989	Y	BR	9
6	Dallas B	6	2,002	N	BR	9
13	Dallas C	6	1,996	N	BR	9
8	Houston A	11	1,998	Y	BR	9
5	Houston B	11	2,005	N	BR	9
19	Houston C	11	1,964	N	BR	9
1	Los Angeles A	14	1,956	Y	BR	9
41	Los Angeles B	14	1,969	N	BR	9
22	San Antonio A	20	1,963	Y	BR	9

BRANCHTYPE	REPBRANCH_SUM
BR	927
HO	<Null>
RE	14

CS Branch MGR

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	BRANCHID	INTEGER		INT				Branch Code
2	EMPCODE	NVARCHAR	8	STRING				Employee ID
3	MGRSINCE	INTEGER		INT				Branch Manager Start Date
4	MGR_NO	INTEGER		INT				Appointment No. of branch man...

Total Number of Records:48

No Primary Key

No Missing Values

No Negative Values

CS Card Transactions

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	KxIndex	BIGINT		FIXED	X(1)	X		
2	SHOP	NVARCHAR	255	STRING				
3	PRODUCT	BIGINT		FIXED				Product
4	MATERIAL	BIGINT		FIXED				Card Material
5	CARD	BIGINT		FIXED				Card Identifier
6	QUANTITY	BIGINT		FIXED				Quantity
7	AMOUNT	BIGINT		FIXED				Transaction Amount
8	TXN_DATE	TIMESTAMP		LONGDATE				
9	DISCOUNT	BIGINT		FIXED				Card Discount

Total Number of Records:48

Primary Key is KxIndex

No Missing Values

No Negative Values

Start Date is from 12 Aug 2012 to 17 May 2015

Highest number of SHOP is MS4 which is having highest number of sales and quantity too.

SHOP	TXN_DATE_COUNT	QUANTITY_SUM	AMOUNT_SUM
MS4	497,467	538,837	327,720,366
M07	477,060	513,929	324,606,017
MB3	362,296	394,007	214,634,719
M03	283,042	312,812	159,230,097
M46	262,890	283,449	150,877,825
M02	249,793	283,776	127,849,644
M23	198,163	222,380	98,238,531
MM4	110,040	118,911	86,081,725
ML2	111,046	119,450	75,653,318
M35	138,995	156,534	69,419,079
MA5	97,428	108,897	61,205,114
MP3	99,809	110,403	60,467,766
M06	102,442	113,969	53,405,147
ME2	91,008	105,942	49,332,202
ML5	57,113	61,447	45,578,766
M18	71,879	82,233	34,563,364
MV6	20,171	22,091	12,320,142
MV3	15,298	16,791	9,685,352
MX7	301	302	263,176

CS City

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	CITYID	INTEGER		INT	X(1)	X		City ID
2	CITY	NVARCHAR	20	STRING				City Name
3	REGION	NVARCHAR	20	STRING				Region
4	POPULATION	INTEGER		INT				Population
5	LATITUDE	DECIMAL	10,5	FIXED				
6	LONGITUDE	DECIMAL	10,5	FIXED				

Total Number of Records:25

Primary Key is CITYID

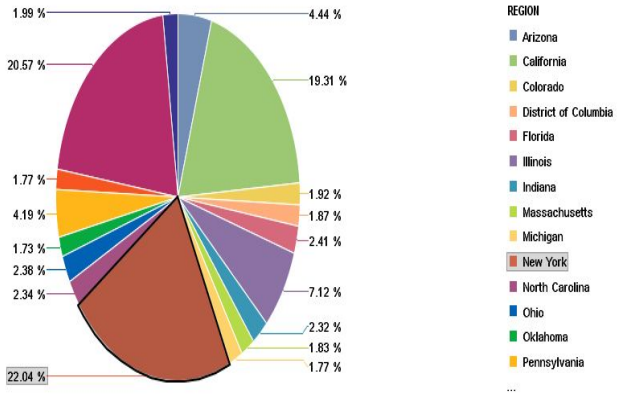
No Missing Values

No Negative Values

Most Number of Branches in REGION is Texas-6

Highest Population is in New York with 22.04%

REGION	CITY_COUNT
Texas	6
California	4
Arizona	1
Colorado	1
District of Columbia	1
Florida	1
Illinois	1
Indiana	1
Massachusetts	1
Michigan	1
New York	1
North Carolina	1
Ohio	1
Oklahoma	1
Pennsylvania	1
Tennessee	1
Washington	1



CS Client

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	CLIENT_ID	INTEGER		INT	X(1)	X		Client/Customer ID
2	HOME_GEOCODE_ID	INTEGER		INT				
3	GENDER	NVARCHAR	1	STRING				Gender
4	BIRTH_DATE	DATE		DAYDATE				Birth Date
5	BRANCHID	INTEGER		INT				Branch Identifier
6	R	INTEGER		INT				Y

Total Number of Records:26,845

Primary Key :CLIENTID

Foreign Key :HOME_GEOCODE_ID ,BRANCHID

No Missing Values

No Negative Values

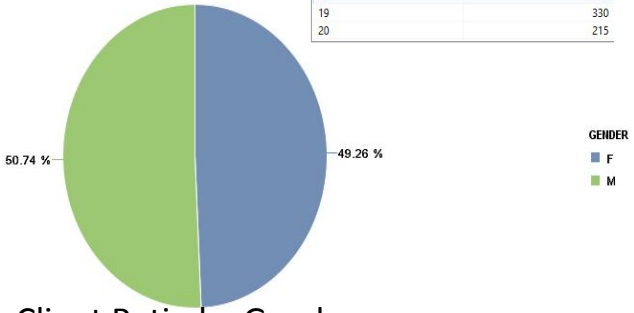
Start Date from 08 Dec 1926 to 21 Dec 1987

Most Number of Accounts in HOME_GEOCODE_ID-1

Most Number of Accounts in BRANCH_ID-7

BRANCHID	BIRTH_DATE_COUNT
7	2,132
42	2,089
10	2,041
36	2,015
1	1,534
41	1,496
19	640
49	627
22	624
8	597
16	594
31	587
2	576
5	554
48	553
38	538
29	524
18	517
3	507
4	449

HOME_GEOCODE_ID	BIRTH_DATE_COUNT
1	3,315
2	230
3	315
4	250
5	355
6	265
7	225
8	345
9	300
10	270
11	280
12	220
13	265
14	225
15	315
16	305
17	270
18	300
19	330
20	215



Client Ratio by Gender

CS Client Contact

CS_CLIENT_CONTACT | TRAINING

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null
1	ID	INTEGER		INT	X(1)	X
2	CLIENT_ID	INTEGER		INT		
3	CONTACTTYPE	NVARCHAR	10	STRING		
4	CONTACT_DETAILS	NVARCHAR	500	STRING		
5	PREFERRED	NVARCHAR	1	STRING		

Total records: 50,984
Primary Key: ID
No missing values
No Negative values
No Account date range.

CS Contact Type

CS_CONTACTTYPE | TRAINING

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null
1	CONTACTTYPE	NVARCHAR	10	STRING	X(1)	X

Total records:3
Primary Key: Contact Type
No missing values
No Negative values
Account date: NA

CS Credit Card

CS_CREDIT_CARD | TRAINING | Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default	Comment
1	CARD_ID	INTEGER		INT	X(1)	X		Card Identifier
2	DISP_ID	INTEGER		INT	X(2)	X		
3	CARD_TYPE	NVARCHAR	7	STRING				Card Type / Product
4	CARD_DATE	DATE		DAYDATE				Card Issue Date

Total records: 4,460
Primary Key: CARD_ID, DISP_ID
No missing values
No Negative values
Account date from 17 Jan 2016 to 13 Jan 2018

CS Dept

CS_DEPT | TRAINING | Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default	Comment
1	DEPT	NVARCHAR	2	STRING				Department ID
2	CENTRALFN	NVARCHAR	1	STRING				Central/HO Function

Total records: 7
Primary Key: NA
No missing values
No Negative values
Account date: NA

CS Dept text

CS_DEPTEXT | TRAINING | Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default	Comment
1	DEPT	NVARCHAR	2	STRING	X(1)	X		Department ID
2	NAME	NVARCHAR	40	STRING				Department Name
3	LANG	NVARCHAR	1	STRING	X(2)	X		Language

Total records: 28
Primary Key: Dept & Lang
No missing values
No Negative values
Account date: NA

CS Designation

CS_DESIGNATION | TRAINING | Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default	Comment
1	DESIG	NVARCHAR	8	STRING				Designation ID
2	EXPMIN	DECIMAL	5,1	FIXED				Minimum Experience
3	EXPMAX	DECIMAL	5,1	FIXED				Maximum Experience
4	BANDS	INTEGER		INT				
5	SALMIN	DECIMAL	10,2	FIXED				
6	SALMAX	DECIMAL	10,2	FIXED				

Total records: 7
Primary Key: NA
No missing values
No Negative values
Account date: NA

CS Designation Text

Total Record: 28

Defined 7 Designation roles:

- **Intern**
- **Junior**
- **Senior**
- **Team Lead**
- **Principal Lead Manager**
- **Function Head**
- **Boardroom Manager**

Language spoken:

English, Mandarin, French and Spanish

NOTE: Primary Key- Designation Name

No Missing Data, no negative value

DESIG	EXPMIN	EXPMAX	BANDS	SALMIN	SALMAX
BOARDM	15	35	1	300,000	1,000,000
FNHEAD	14	35	1	195,000	450,000
INTERN	0	3	1	18,000	36,000
JUNIOR	1	5	2	24,000	65,000
PLTMGR	10	25	2	140,000	270,000
SENIOR	4	9	3	42,000	110,000
TELEAD	7	15	2	90,000	180,000

CS EMP_CTC

Table Name: CS_EMP_CTC

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default	Comment
1	EMPCODE	NVARCHAR	8	STRING				
2	CURR_SALARY	DECIMAL	12	FIXED				

Total Staff records	117
Primary Key	NA
No Missing Values	
No Negative values	

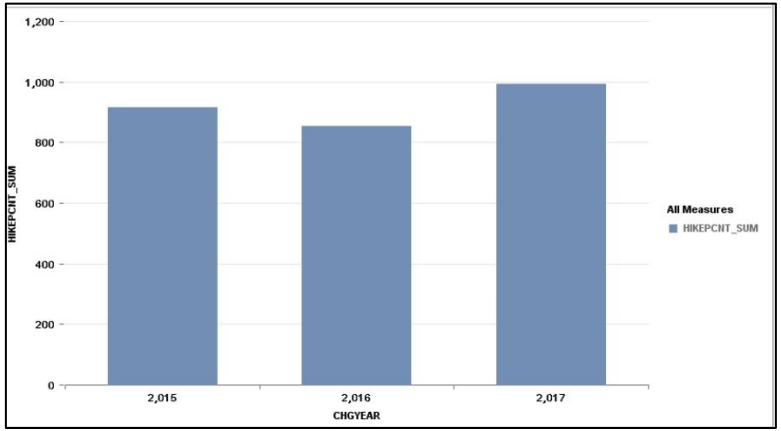
CS EMP_SAL_CHANGE

Table Name: CS_EMP_SALCHANGE | Schema: TRAINING | Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default	Comment
1	EMPCODE	NVARCHAR	8	STRING				
2	CHGYEAR	INTEGER		INT				
3	HIKEPCNT	DECIMAL	5,3	FIXED				

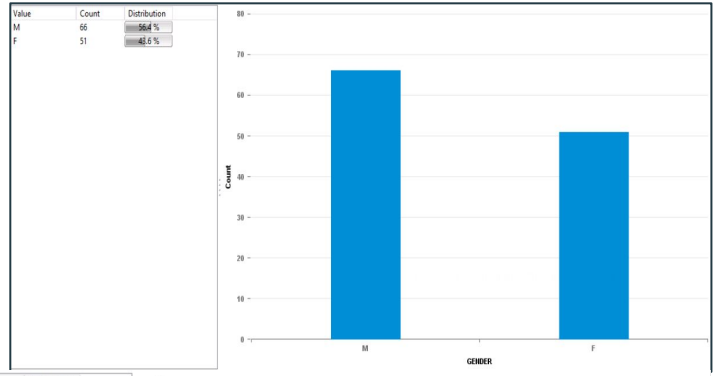
Account Period (Year)	2015-2017
Total records	200
Staff employed (Constant)	117
Staff Turnover (2015-2017)	83
Primary Key	Employee Code
Any Missing data	0
Any Negative Values	0



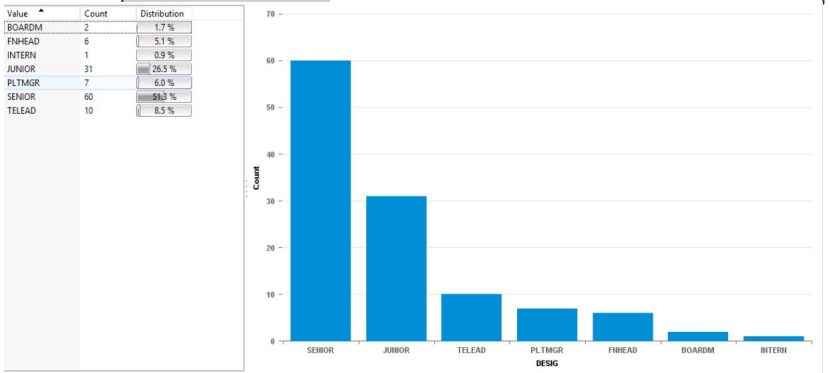
CS Employee

Total Staff Strength	117
Male/Female	51/66
Staff Turnover	83

Total records: 117
Primary Key: EMPCODE_ID
Foreign Key: CITYID
Missing values: 61
No Negative values
DOB Range: 1962-1996



Value	Count	Distribution
M	66	56.4 %
F	51	43.6 %



Value	Count	Distribution
BOARDM	2	1.7 %
FNHEAD	6	5.1 %
INTERN	1	0.9 %
JUNIOR	31	26.5 %
PLTMGR	7	6.0 %
SENIOR	60	51.3 %
TELEAD	10	8.5 %

CS_GEODEMO

Table Name: CS_GEODEMO

Schema: TRAINING

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default	Comment
1	GEOCODE_ID	INTEGER		INT	X(1)	X		
10	URBAN_RATIO	DECIMAL	7,3	FIXED				
11	AVERAGE_SALARY	INTEGER		INT				
12	UNEMP_RATE_2016	DECIMAL	6,2	FIXED				
13	UNEMP_RATE_2017	DECIMAL	6,2	FIXED				
14	RATIO_ENTREPRENEURS	INTEGER		INT				
15	NUMBER_CRIMES_2016	INTEGER		INT				
16	NUMBER_CRIMES_2017	INTEGER		INT				
2	DISTRICT_NAME	NVARCHAR	21	STRING				District Name
3	REGION_NAME	NVARCHAR	17	STRING				Region Name
4	NUMBER_INHABITANTS	INTEGER		INT				Number of Inhabitants
5	NUMBER_CITIES_S	INTEGER		INT				
6	NUMBER_CITIES_M	INTEGER		INT				
7	NUMBER_CITIES_L	INTEGER		INT				
8	NUMBER_CITIES_XL	INTEGER		INT				
9	NUMBER_CITIES	INTEGER		INT				

Total Records: 77
Primary Key: GEOCODE_ID
No missing values
No Negative values

Table shows there are 77 distinct district names and 8 distinct region names, and the total no of inhabitants is 10,309,137.

CS GL Entry

We cannot find any raw data in this table.

Table Name: CS_GLENTRY

Schema: TRAINING

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Dimension	Column Store Data Type	Key	Not Null	Default
1	ACCTRANSID	INTEGER		INT			
2	LINEID	INTEGER		INT			
3	TXN_DATE	DATE		DAYDATE			
4	SOURCE_TRANSID	INTEGER		INT			
5	GLACCOUNT	INTEGER		INT			
6	DRCR	NVARCHAR	1	STRING			
7	AMOUNT	DECIMAL	12,2	FIXED			
8	CURRENCY	NVARCHAR	3	STRING			

Table Name: CS_GL_CONFIG

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	ACCOUNT_TYPE	NVARCHAR	10	STRING	X(1)	X	
2	TRANSMODE	NVARCHAR	2	STRING	X(2)	X	
3	DRCR	NVARCHAR	1	STRING	X(3)	X	
4	GLDRCR	NVARCHAR	1	STRING			
5	LINE_NUM	TINYINT		INT	X(4)	X	
6	GLCODE	INTEGER		INT			

Total records : 42
Primary key: ACCOUNT_TYPE, TRANSMODE, DRCR, LINE_NUM
No Missing Value
No negative values
No Date range

Table Name: CS_LOAN

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	ID	INTEGER		INT	X(1)	X	
2	ACCOUNT_ID	INTEGER		INT			
3	DURATION	INTEGER		INT			
4	STATUS	NVARCHAR	1	STRING			
5	AMOUNT	INTEGER		INT			
6	PAYMENT	INTEGER		INT			
7	DATE	DATE		DAYDATE			

Total records : 3,410
Primary key: ID
No Missing Value
No negative values
Jun 7, 2013 to Dec 8, 2017

Table Name:	Schema:	Type:
CS_LOYALTY_CARD	TRAINING	Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	CARD	BIGINT		FIXED	X(1)	X	
2	GENDER_TYPE	NVARCHAR	255	STRING			
3	AGE_GROUP	NVARCHAR	255	STRING			
4	EMAIL_FLG	NVARCHAR	255	STRING			
5	ADDRESS_FLG	NVARCHAR	255	STRING			
6	LOYALTY_START_DATE	TIMESTAMP		LONGDATE			

Total records : 328,106
Primary key: Card
There are missing value
No negative values
Apr 10, 2004 to Nov 23, 2014

2 distinct values for EMAIL_FLG

Value	Count	Distribution
?	114,811	35.0 %
have email address	213,295	65.0 %

2 distinct values for ADDRESS_FLG

Value	Count	Distribution
?	49,262	15.0 %
have physical ad...	278,844	85.0 %

7 distinct values for AGE_GROUP

Value	Count	Distribution
-	19,420	5.9 %
18 - 25	11,506	3.5 %
26 - 35	89,204	27.2 %
36 - 45	88,536	27.0 %
46 - 60	65,115	19.8 %
?? 17	31,485	9.6 %
????? 60	22,840	7.0 %

Table Name: CS_MAPPING_TABLE

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Defau
1	DISP_ID	INTEGER		INT	X(1)	X	
2	CLIENT_ID	INTEGER		INT			
3	ACCOUNT_ID	INTEGER		INT			
4	CLIENT_TYPE	NVARCHAR	5	STRING			

Total records : 26,845
Primary key: DISP_ID
There are no missing values
No negative values
No Date

Table Name: CS_MODE

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	MODE	NVARCHAR	2	STRING			
2	DESCRIPTION	NVARCHAR	40	STRING			
3	EFFECT	NVARCHAR	3	STRING			
4	INITIATOR	NVARCHAR	3	STRING			

Total records : 8
Primary key:
There are no missing values
No negative values
No Date

Table Name: CS_PRODUCT

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	PRODUCTCODE	INTEGER		INT			
2	PRODUCTNAME	NVARCHAR	20	STRING			
3	PGID	NVARCHAR	4	STRING			
4	REVAL	NVARCHAR	1	STRING			

Total records : 7
Primary key:
There are no missing values
No negative values
No Date

Table Name: CS_PRODUCTGROUP

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	PGID	NVARCHAR	4	STRING	X(1)	X	
2	PGNAME	NVARCHAR	20	STRING			
3	ASSETLIAB	NVARCHAR	1	STRING			

Total records : 4
Primary key: PGID
There are no missing values
No negative values
No Date

Table Name: CS_TRANSACTIONS

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	TXN_ID	BIGINT		FIXED	X(1)	X	
2	ACCOUNT_ID	INTEGER		INT			
3	CREDIT	NVARCHAR	1	STRING			
4	MODE	NVARCHAR	2	STRING			
5	TYPE	NVARCHAR	1	STRING			
6	BANK	NVARCHAR	5	STRING			
7	ACCOUNT	INTEGER		INT			
8	AMOUNT	DECIMAL	11,5	FIXED			
9	BALANCE	DECIMAL	14,6	FIXED			
10	TXN_DATE	DATE		DAYDATE			

8 distinct values for TYPE

Value	Count	Distribution
?	2,676,570	50.7 %
C	915,570	17.3 %
F	779,160	14.8 %
H	590,325	11.2 %
P	151,690	2.9 %
I	92,500	1.8 %
L	67,900	1.3 %
N	7,885	0.1 %

6 distinct values for MODE

Value	Count	Distribution
D1	2,174,590	41.2 %
D2	1,041,415	19.7 %
?	915,570	17.3 %
C1	783,715	14.8 %
C2	326,130	6.2 %
D3	40,180	0.8 %

Total records : 5,281,600
Primary key: TXN_ID
Foreign Key: ACCOUNT_ID
There are missing values
There are negative values
Oct 28, 2014 to Feb 28, 2017

Table Name: CS_TRANSACTIONS_BACKEND

Schema: TRAINING

Type: Column Store

Columns | Indexes | Further Properties | Runtime Information

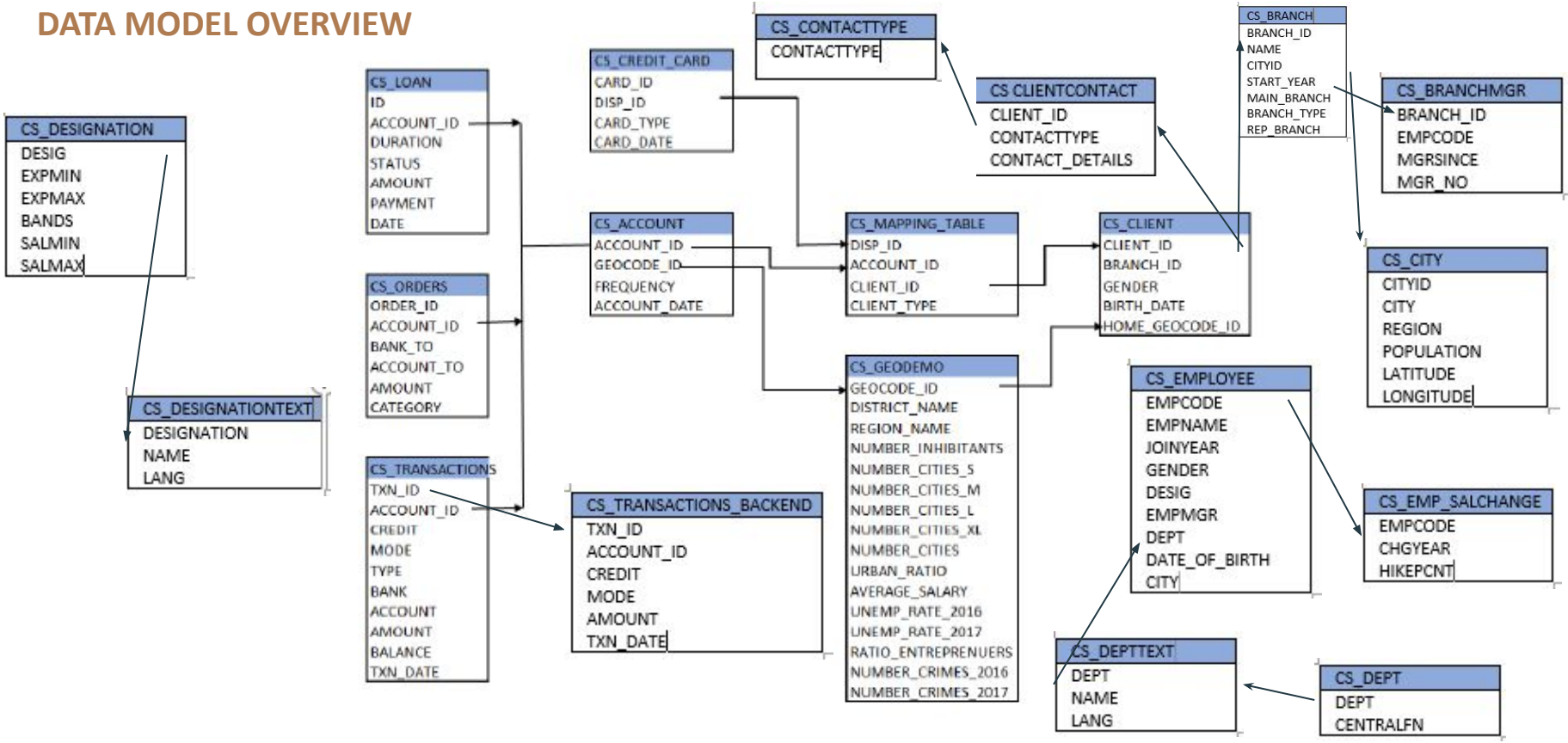
	Name	SQL Data Type	Di...	Column Store Data Type	Key	Not Null	Default
1	TXN_ID	BIGINT		FIXED	X(1)	X	
2	ACCOUNT_ID	INTEGER		INT			
3	CREDIT	NVARCHAR	1	STRING			
4	MODE	NVARCHAR	2	STRING			
5	AMOUNT	DECIMAL	11,5	FIXED			
6	TXN_DATE	DATE		DAYDATE			

200 distinct values for AMOUNT

Value	Count	Distribution
?	75,268	7.8 %
3.65	2,373	0.2 %
5	10,316	1.1 %
5.6	5	0.0 %

Total records : 965,306
Primary key: TXN_ID
There are missing values
There are no negative values
Jan 31, 2012 to Jan 31, 2018

DATA MODEL OVERVIEW



Data Preparation

Data Preparation Tasks

Select data

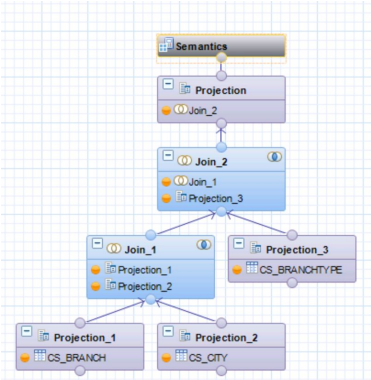
Clean data

Construct data

Integrate data

Format data

DIM Branch



Key	Name	Label
☐	12 BRANCHID	Branch Identifier
☐	AB NAME	Branch Name
☐	12 STARTYEAR	Branch Start Date
☐	AB MAINBRANCH	City/Region Main Branch
☐	AB BRANCHTYPE	Level/Type of Branch
☐	12 REPBRANCH	Reporting / Managing Branch
☐	AB CITY	City Name
☐	AB REGION	Region
☐	AB BRANCH_TYPE	Branch Type Description

	Branch		
Table	CS_BRANCH		
Joins	NA	CS_BRANCHTYPE	CS_CITY
Cardinality	NA	NA	NA
Calculated Column	NA	NA	NA
Output	Branch ID, Start Year, Main Branch, Reporting Branch, Region, City ID	Branch ID, Branch Name, Branch Type	City ID, City Name

	Title	Formula	Purpose
Calculated Column	NA	NA	NA

DIM Branch Organisation Structure

48 rows retrieved - 31 ms

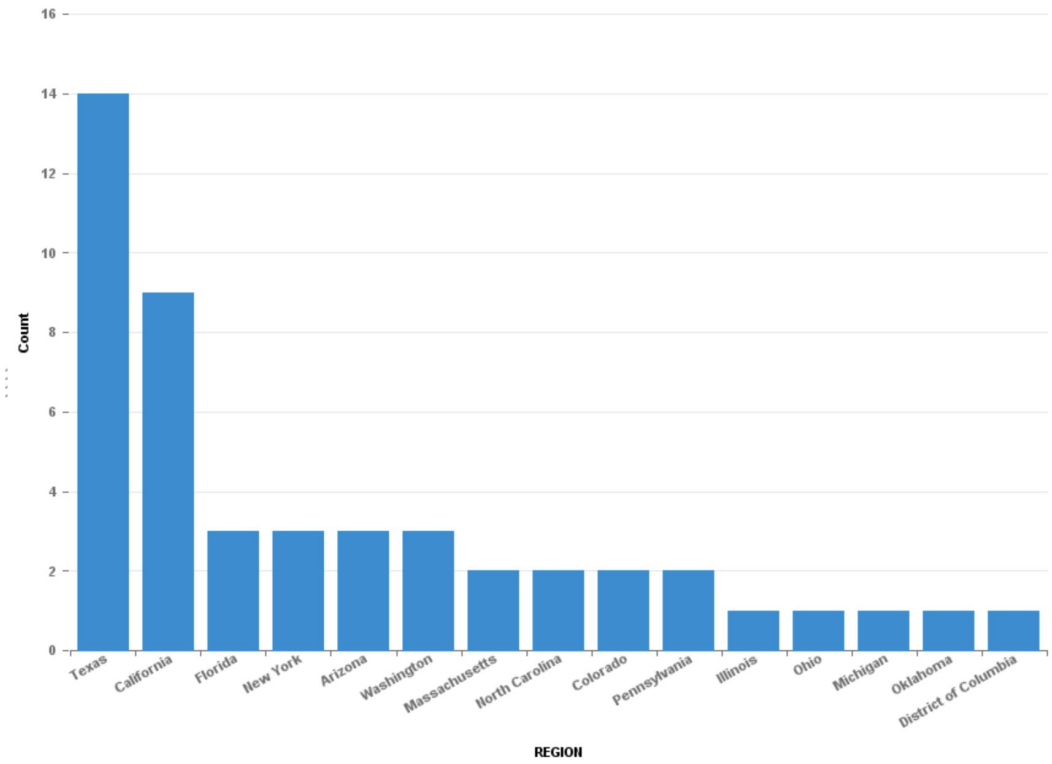
Chart | Table | Grid | HTML

BRANCHID	CITY	REGION	BRANCH_TYPE
9	San Franci...	California	Regional Centre
36	Chicago	Illinois	Regional Centre
7	New York	New York	Head Office
1	Los Angeles	California	Branch
2	Philadelphia	Pennsylvania	Branch
3	Detroit	Michigan	Branch
4	Phoenix	Arizona	Branch
5	Houston	Texas	Branch
6	Dallas	Texas	Branch
8	Houston	Texas	Branch
10	New York	New York	Branch
11	Seattle	Washington	Branch
12	Austin	Texas	Branch
13	Dallas	Texas	Branch
14	San Jose	California	Branch
15	Denver	Colorado	Branch
16	San Antonio	Texas	Branch
17	Seattle	Washington	Branch
18	Oklahoma ...	Oklahoma	Branch
19	Houston	Texas	Branch
20	Phoenix	Arizona	Branch

BRANCH_TYPE	BRANCHID_COUNT
Branch	45
Head Office	1
Regional Centre	2

There are 45 branches, 1 Head Office and 2 Regional Centres

DIM Branch Organisation Structure



15 distinct values for REGION

Value	Count	Distribution
Texas	14	29.2 %
California	9	18.8 %
Florida	3	6.3 %
New York	3	6.3 %
Arizona	3	6.3 %
Washington	3	6.3 %
Massachusetts	2	4.2 %
North Carolina	2	4.2 %
Colorado	2	4.2 %
Pennsylvania	2	4.2 %
Illinois	1	2.1 %
Ohio	1	2.1 %
Michigan	1	2.1 %
Oklahoma	1	2.1 %
District of Colum...	1	2.1 %

Both Texas and California have the most branches.

DIM Branch Organisation Structure

48 rows retrieved - 32 ms

Chart | Table | Grid | HTML

REGION	NAME	MAINBRANCH
Arizona	Phoenix A	Y
California	Los Angeles A	Y
California	San Diego A	Y
California	San Francisco A	Y
California	San Jose A	Y
Colorado	Denver A	Y
Florida	Jacksonville A	Y
Massachusetts	Boston A	Y
New York	New York A	Y
North Carolina	Charlotte A	Y
Pennsylvania	Philadelphia A	Y
Texas	Austin A	Y
Texas	Dallas A	Y
Texas	El Paso A	Y
Texas	Houston A	Y
Texas	San Antonio A	Y
Washington	Seattle A	Y
Arizona	Phoenix B	N
Arizona	Phoenix C	N
California	Los Angeles B	N
California	San Diego B	N

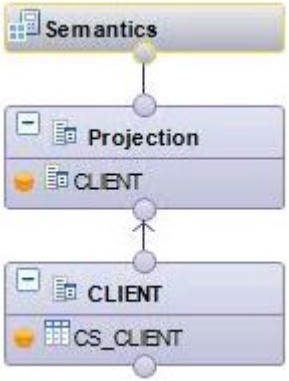
The following regions have main branches:

1) Arizona (Phoenix A)
2) California
3) Colorado
4) Florida
5) Massachusetts
6) New York
7) North Carolina
8) Pennsylvania
9) Texas
10) Washington

2 distinct values for MAINBRANCH

Value	Count	Distribution
N	31	64.6 %
Y	17	35.4 %

DIM Customer

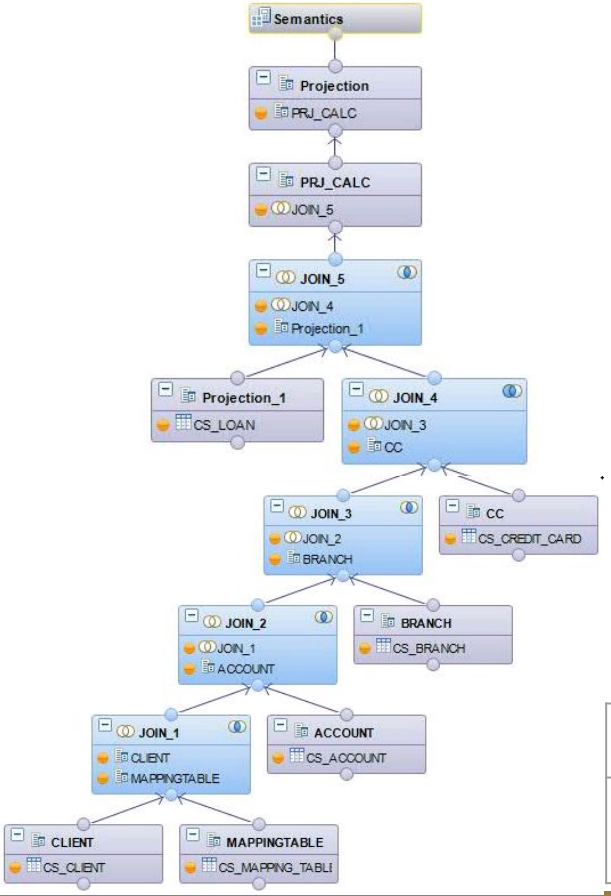


	Customer
Table	CS_CLIENT
Joins	NA
Cardinality	NA
Calculated Column	NA
Output	Client ID, Gender, Birth date, Branch ID

Statistics for [CLIENT_ID]	
Data Type	Numeric
Native datatype	INTEGER
Java type	java.lang.Integer
Column display size	10
Column precision	10
Column scale	0
Row Count	26,845
Min value	1
Max value	69,998
Average	31,359.012
Sum	841,832,675

	Title	Formula	Purpose
Calculated Column	NA	NA	NA

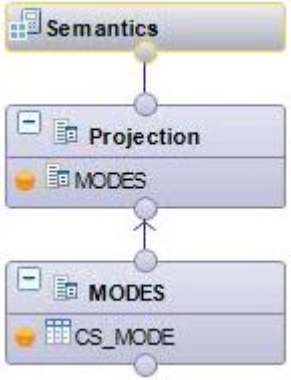
DIM Customer A/C



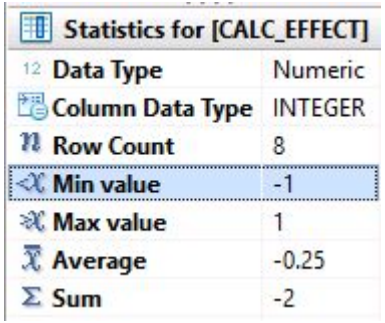
	JOIN_1	JOIN_2	JOIN_3	JOIN_4	JOIN_5	Projected Calculation
Table	CS_CLIENT CS_MAPPING_TABLE	JOIN_1 CS_ACC	JOIN_2 CS_BRANCH	JOIN_3 CS_CREDIT_CARD	JOIN_4 CS_LOAN	JOIN_5
Joins	Inner	Inner	Inner	Left Outer	Inner	NA
Cardinality	1:N	1:N	N:1	N:M	1:1	
Calculated Column	NA	NA	NA	NA	NA	ACCOUNT_TYPE
Output	All Clients	All Clients & ACC	All Clients, ACC & Branches	All Clients, ACC, Branches & Credit Cards	All Clients, ACC, Branches & Credit Cards	NA

	Title	Expression	Purpose
Calculated Column	ACCOUNT_TYPE	IF(ISNULL("CARD_TYPE"), IF(ISNULL("LOAN_ID"), 'CHECKING', 'LOAN'), 'CREDITCARD')	Identify different type of transactions from Checking, Loans or Credit Card.

DIM Trans Mode



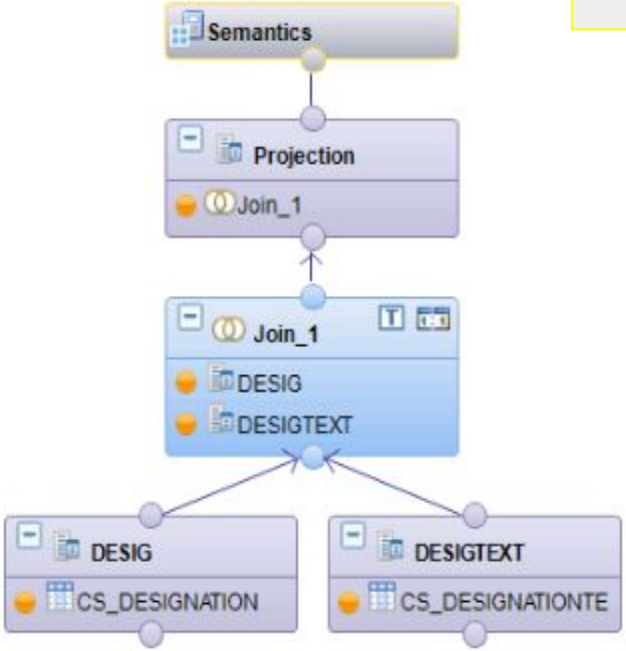
	Trans Mode
Table	CS_MODE
Joins	NA
Cardinality	NA
Calculated Column	NA
Output	Mode, Description, Effect, Initiator, Calc Effect



Statistics for [CALC_EFFECT]	
Data Type	Numeric
Column Data Type	INTEGER
Row Count	8
Min value	-1
Max value	1
Average	-0.25
Sum	-2

	Title	Expression	Purpose
Calculated Column	DIM_TRANS _MODES	IF ("EFFECT"='INC', 1, -1)	times +1 is increasing effect while times -1 will give a negative value indicates decreasing

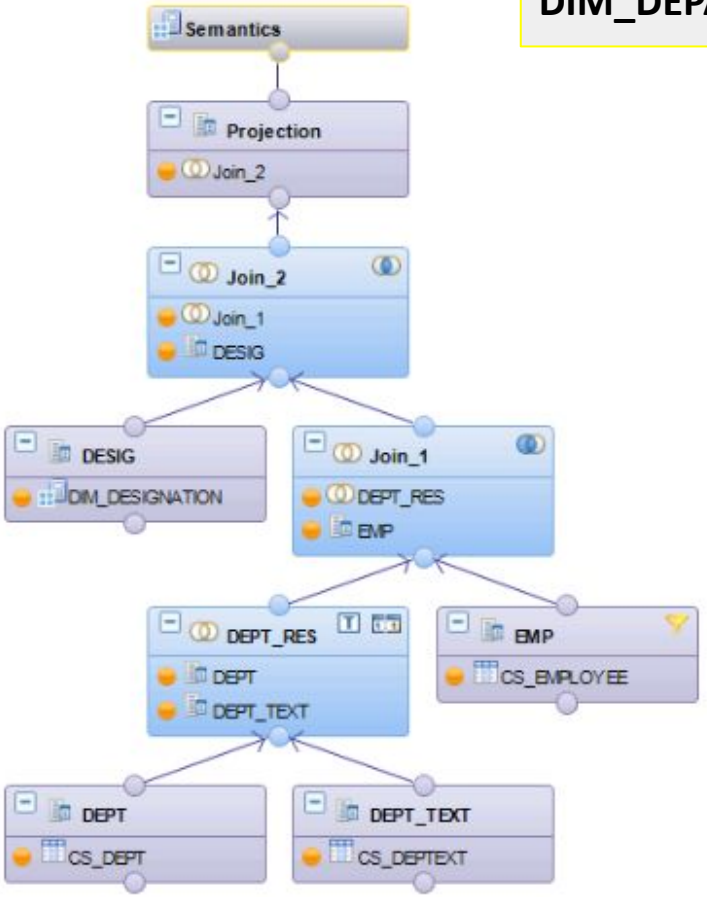
DIM_DESIGNATION



		JOIN 1
Table	CS_DESIGNATION CS_DESIGNATIONTEXT	
Joins		Text Joins
Cardinality		1..1
Output		DESIG (Designation ID) DESCRIPTION (Designation roles) LANGUAGE

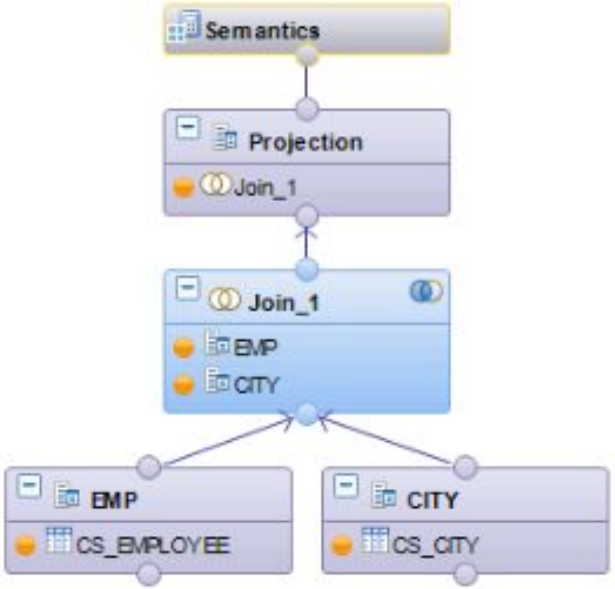
7 Designation roles, language in English

DIM_DEPARTMENT



	DEPT_RES	JOIN_1	JOIN_2
Table	CS_DEPT (DEPT) CS_DEPTEXT(DEPT_TEXT)	DEPT_RES CS_EMPLOYEE(EMP)	JOIN_1 DESIG(DIM_DESIGNATION)
JOINS	TEXT JOINS	LEFT OUTER	INNER
Output	DEPT (DEPT ID) DEPT NAME LANG CENTRALFN	EMPCODE DESIG FUNCTIONHEAD CENTRALFN LANG DEPT	EMPCODE DESIG FUNCTIONHEAD CENTRALFN DEPT NAME LANG DEPT DESIGNATION

DIM_EMPLOYEE



	JOIN_1
TABLE	CS_EMPLOYEE (EMP) CS_CITY(CITY)
JOINS	LEFT OUTER
OUTPUT	EMPCODE,EMPNAME,JOINYEAR,GENDER,DOB,DEPT,CITY,AGE, TITLE,DESIG,EMPMGR,

Missing value in City (61)
Gender:F(51) M(66)
Total staff :117

VW_HR_EMPLOYEE

	JOIN_1
TABLES	CS_EMPLOYEE DIM_EMPLOYEE
CALCULATED COLUMNS	EMPLOYEE_COUNT EMPLOYEE_DOB EMPLOYEE_AGE IS_BIRTH_MONTH
HIERARCHY	EMPLOYEE-MANAGER
OUTPUT	ALL EMPLOYEES

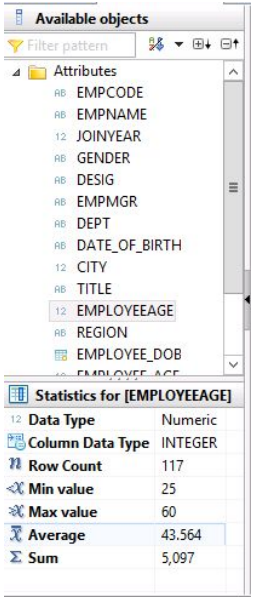
CALCULATED COLUMN	EMPLOYEE_COUNT	1	TO REPRESENT NUMBER OF EMPLOYEES
CALCULATED COLUMN	EMPLOYEE_DOB	to_date(“DATE_OF_BIRTH”)	TO CONVERT THE DATE OF BIRTH IN YYYYMMDD FORMAT
CALCULATED COLUMN	EMPLOYEE_AGE	days_between(“EMPLOYEE_DOB”, current_date),365.25	TO CALCULATE THE AGE OF EMPLOYEES
CALCULATED COLUMN	IS_BIRTH_MONTH	casewhen month(“DATE_OF_BIRTH”)=month(current_date)then’Y’ELSE’N’END	TO REPRESENT IF THE CURRENT MONTH IS THE BIRTH MONTH OF EMPLOYEE

VW_HR_EMPLOYEE DATA PREVIEW

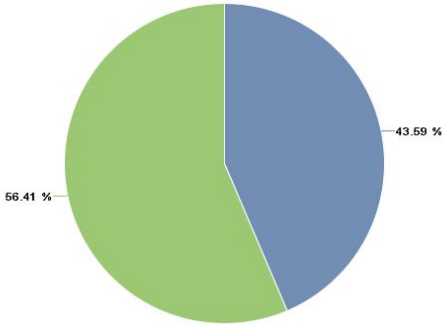
ANALYSIS

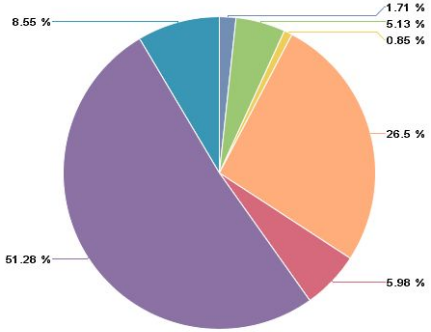
-Total Employee Count is 117
-Percentage of Male to Female is 56.41:43.59
-Most common designation is SENIOR and Least common designation is INTERN
-61 Employees have REGION values null
-Highest Employee Age is 60 and lowest Employee Age is 25

EMPLOYEE_COUNT_SUM
117



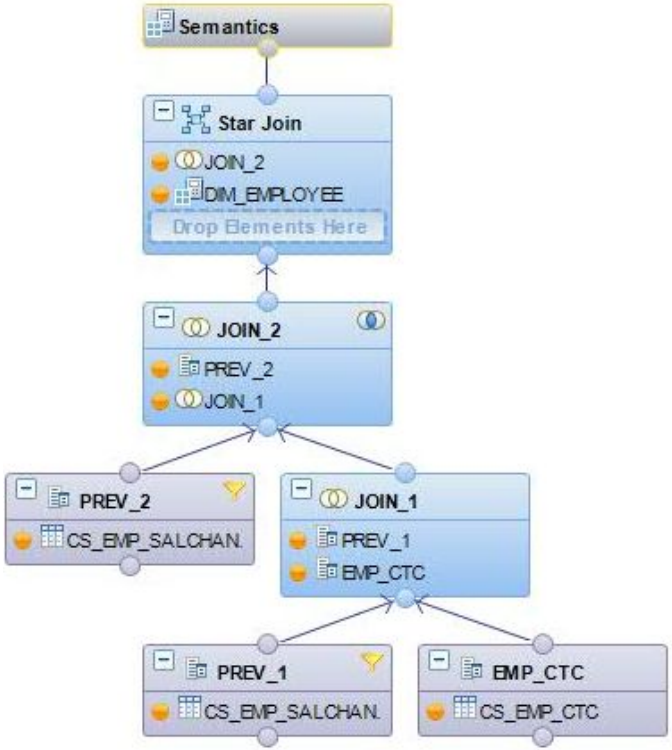
REGION	EMPLOYEE_COUNT_SUM
<Null>	61
Arizona	3
California	9
Colorado	2
District of Colu...	1
Florida	4
Massachusetts	2
Michigan	1
New York	11
North Carolina	2
Ohio	1
Oklahoma	1
Pennsylvania	2
Texas	14
Washington	3





Dan

VW_HR_EMPSAL_V1



	Join 1	Join 2	Star Join
Tables	CS_EMP_SALCHANGE CS_EMP_CTC	Join_1 CS_EMP_SALCHANGE	
Calc columns	PREV_YEAR_SAL	PREV_YEAR2_SAL	
Hierarchy			Child: EMPCODE Parent: EMPMGR
Output	Employee Salary	Employee Salary	Employee Salary

PREV_YEAR1_SAL	**ROUND**("CURR_SALARY"/(1+"HIKEPCNT_P1"/100),0)	To calculate existing salary after the hike
PREV_YEAR2_SAL	**ROUND**("PREV_YEAR1_SAL"/(1+"HIKEPCNT_P2"/100),0)	To calculate salary before the hike for the previous year

VW_HR_EMPSAL_V1

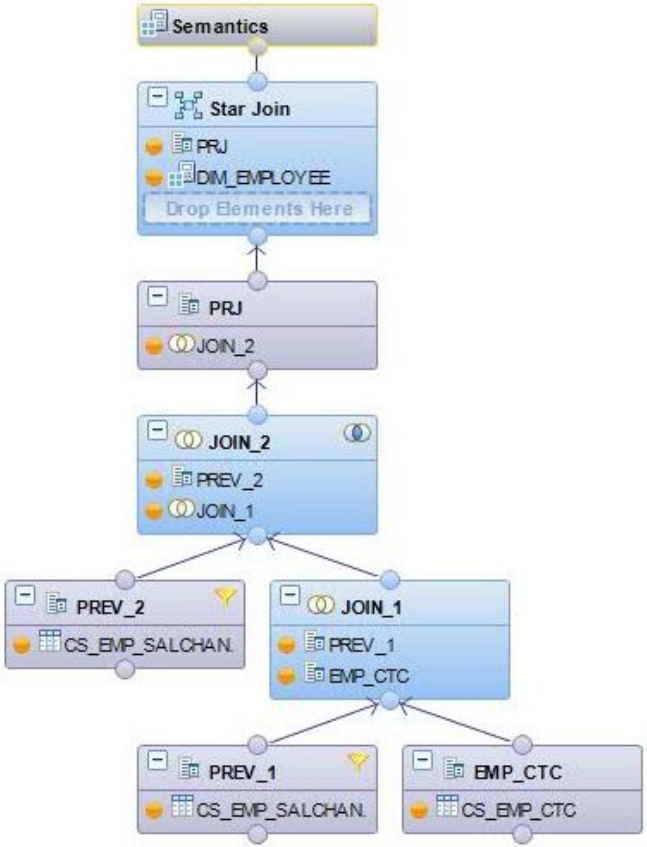
GENDER	HIKEPCNT_P1_SUM	HIKEPCNT_P2_SUM
F	8.713	7.407
M	8.287	7.186

Hike % between the genders are similar.

DEPT	AVERAGE_DIFF_SUM	CURR_SALARY_SUM	PREV_YEAR1_SAL_SUM	HIKEPCNT_P1_SUM
BD	-0.04	1,861,900	1,694,229	9.9
CS	0.79	1,313,300	1,198,643	9.49
FI	17.112	1,603,900	1,466,496	8
FL	2.212	6,751,800	6,228,109	8.226
HR	-8.675	911,300	839,458	9.371
PL	15.067	981,600	890,779	8.86
SL	7.462	1,376,100	1,259,803	8.59

Comparing average hike difference between departments as a whole company.

VW_HR_EMPSAL_V2



	Join 1	Join 2	Star Join
Tables	CS_EMP_SALCHANGE CS_EMP_CTC	Join_1 CS_EMP_SALCHANGE	
Calc columns	PREV_YEAR_SAL	PREV_YEAR2_SAL	
Hierarchy			Child: EMPCODE Parent: EMPMGR
Output	Employee Salary	Employee Salary	Employee Salary
Input Parameter			IP_CURRENT_YEA R IP_TARGET_CURRE NCY

PREV_YEAR1_SAL	**ROUND**("CURR_SALARY"/(1+" HIKEPCNT_P1"/100),0)	To calculate existing salary after the hike
PREV_YEAR2_SAL	**ROUND**("PREV_YEAR1_SAL"/(1+"HIKEPCNT_P2"/100),0)	To calculate salary before the hike for the previous year

VW_HR_EMPSAL_V2

DESIG	HIKEPCNT_P1_SUM	HIKEPCNT_WA1_SUM	HIKEPCNT_P2_SUM	HIKEPCNT_WA2_SUM	AVERAGE_DIFF_SUM
BOARDM	9.9	9.896	7.05	7.051	-0.04
FNHEAD	8.866	8.939	7.783	7.816	0.823
INTERN	9	8.999	6.4	6.398	-0.011
JUNIOR	8.09	8.341	6.619	6.678	3.102
PLTMGR	9.542	9.715	7.314	7.393	1.813
SENIOR	8.13	8.273	7.253	7.375	1.758
TELEAD	10.4	10.491	9.33	9.495	0.875

Comparing weighted average hike % between designations. The smaller the difference, the smaller the deviation between salaries.

VW_HR_EMPSAL_V3

	Join 1	Join 2	Join 3	Join 4	Star Join
Tables	CS_EMP_SALCHANGE CS_EMP_CTC	JOIN_1 CS_EMP_SALCHANGE	JOIN_2 DIM_EMPLOYEE	JOIN_3 AVG_DEPT_DESIG	PRJ_CALC DIM_EMPLOYEE
Calc Columns	PREV_YEAR1_SAL	PREV_YEAR2_SAL		AVG_CAL_SALARY	AVG_LAST_2_YEARS_HIKE MIN_HIKE_AVG_75 HIKE_AVG_FOR_GREATER_THAN_MEDIAN HIKE_AVG_FOR_LESS_THAN_MEDIAN IS_HIGHER_THAN_MEDIAN HIKE_PCNT_RECOMMENDED
Hierarchy					Child: EMPCODE Parent: EMPMGR
Output	Employee Salary	Employee Salary	Desig Dept	Average Current Salary	
Input Parameter	IP_CURRENT_YEAR	IP_CURRENT_YEAR	IP_CURRENT_YEAR	IP_CURRENT_YEAR	IP_CURRENT_YEAR IP_TARGET_CURRENCY

VW_HR_EMPSAL_V3

Chart | Table | Grid | HTML

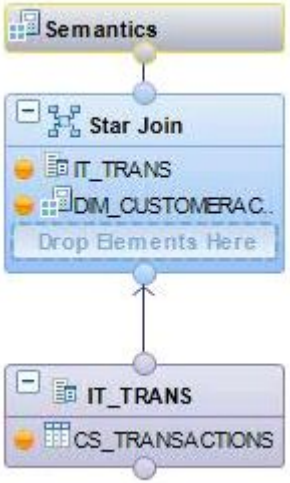
DESIG	EMPCODE		CURR_SALARY_SUM	PREV_YEAR1_SAL_SUM	HIKE_PCNT_RECOMMENDED_SUM
BOARDM	E100010	Y	962,000	876,138	6.337
BOARDM	E100020	N	899,900	818,091	15
FNHEAD	E100220	N	431,300	385,089	15
FNHEAD	E100250	N	298,900	281,185	11.9
FNHEAD	E100340	N	370,000	333,935	15
FNHEAD	E100530	N	403,500	379,229	13.5
FNHEAD	E100540	N	355,000	332,086	15
FNHEAD	E100570	N	347,300	313,448	15
INTERN	E100400	N	26,100	23,945	15
JUNIOR	E100050	Y	59,500	54,637	5.662
JUNIOR	E100060	N	44,900	43,215	10.5
JUNIOR	E100070	N	46,800	44,402	15
JUNIOR	E100080	N	55,400	52,117	14.8
JUNIOR	E100110	N	51,800	47,134	15
JUNIOR	E100140	N	51,300	48,079	13.3
JUNIOR	E100170	N	54,600	51,705	12.5
JUNIOR	E100210	Y	63,200	58,627	5.587
JUNIOR	E100260	N	43,000	40,300	10.4
JUNIOR	E100310	Y	62,600	56,498	7.087
JUNIOR	E100350	Y	61,900	55,367	6.862
JUNIOR	E100390	N	43,400	41,491	9.5

As the bank wants to normalise salaries, the table provides information on what the recommended hike % to be.

VW_HR_EMPSAL_V3

AVG_LAST_2_YEARS_HIKE	(“HIKEPCNT_P1” +”HIKEPCNT_P2”)/2
MIN_HIKE_AVG_75	“AVG_LAST_2_YEARS_HIKE”*0.75
HIKE_AVG_FOR_GREATER_THAN_MEDIAN	LEAST(GREATEST(“MIN_HIKE_AVG_75”,3.0),9.0)
HIKE_AVG_FOR_LESS_THAN_MEDIAN	LEAST(GREATEST(“AVG_LAST_2_YEARS_HIKE”*2.0,3.0),15.0)
IS_HIGHER_THAN_MEDIAN	IF(“CURR_SALARY”>”AVERAGE_CURR_SALARY”,’Y’,’N’)
HIKE_PCNT_RECOMMENDED	IF(‘IS_HIGHER_THAN_MEDIAN’=’Y’,”HIKE_AVG_FOR_GREATER_THAN_MEDIAN”,”HIKE_AVG_FOR_LESS_THAN_MEDIAN”)

VW_TRANSACTIONS_V0



	VW_TRANSACTIONS_V0
TABLES	CS_TRANSACTIONS DIM_CUSTOMERACCOUNT
CALCULATED COLUMNS	-
FILTER	ACCOUNT_ID=1
OUTPUT	Transactions of Bank Customers

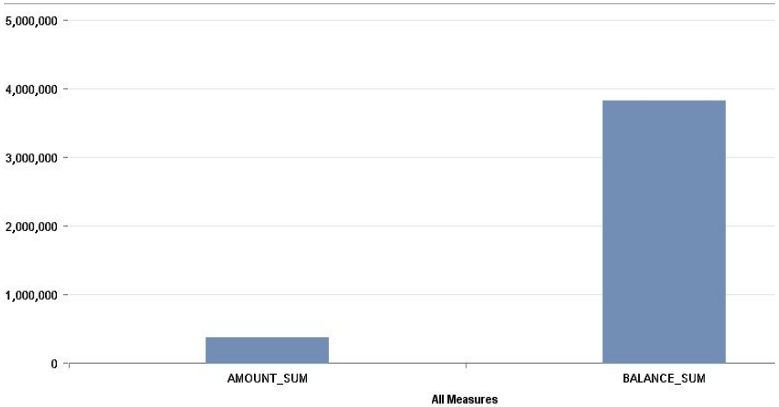
DATA PREVIEW FOR VW_TRANSACTIONS_V0

ANALYSIS

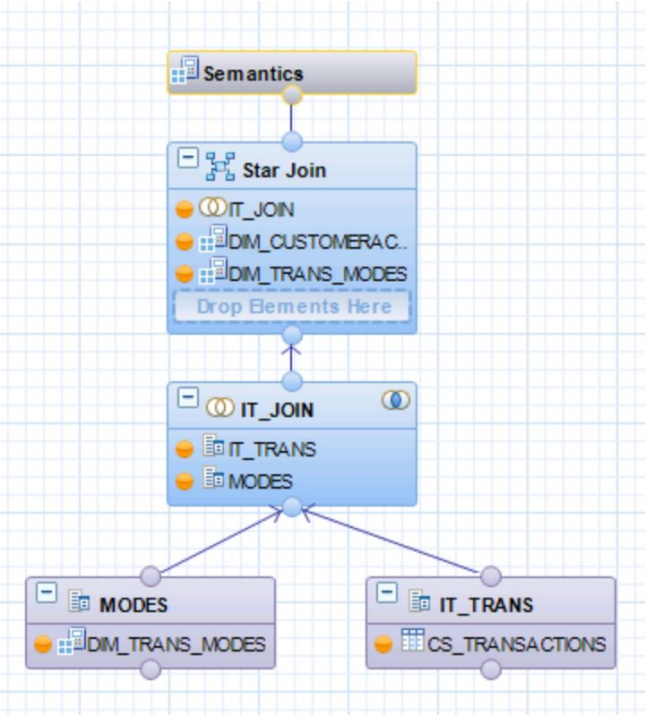
FILTER:ACCOUNTID=1

TXN_DATE	AMOUNT_SUM	BALANCE_SUM
2014-03-24	1,000	1,000
2014-04-13	3,679	4,679
2014-04-23	12,600	17,279
2014-04-30	19.2	17,298.2
2014-05-13	3,679	20,977.2
2014-05-23	2,100	23,077.2
2014-05-31	79	23,156.2
2014-06-13	3,679	26,835.2
2014-06-22	200	26,635.2
2014-06-30	100.6	26,735.8
2014-07-13	3,679	30,414.8
2014-07-22	5,300	25,114.8
2014-07-31	108.9	25,223.7
2014-08-13	3,679	28,902.7
2014-08-21	7,500	21,402.7
2014-08-31	113.8	42,989.2
2014-09-05	2,452	19,035.3
2014-09-13	3,679	22,714.3
2014-09-20	700	22,014.3
2014-09-30	106.1	44,197

- BALANCE_SUM>AMOUNT_SUM
- The effect of Amount is updated in Balance Column after each transactions
- Each row in BALANCE Column sums up values of all preceding rows in AMOUNT Column
- BALANCE should be of type Dimension inspite of being a Measure

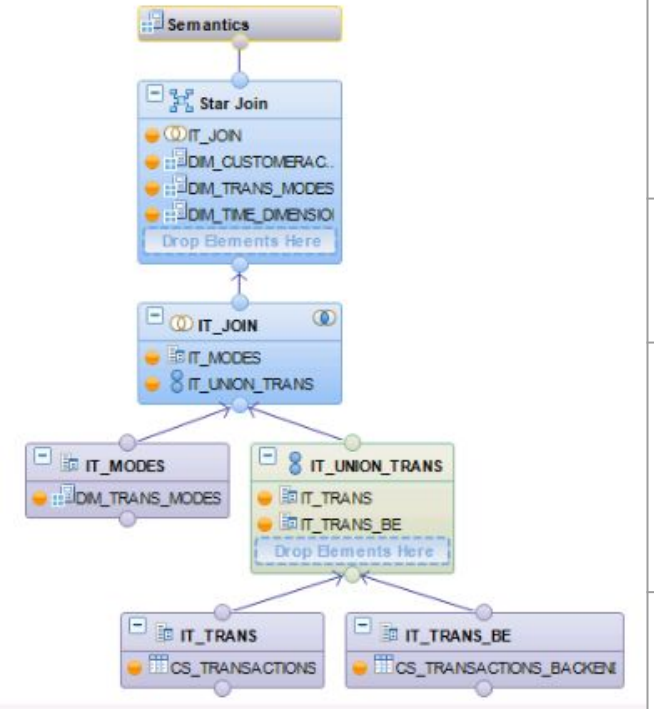


VW_TRANSACTIONS_V1



	Projection_1 (Modes)	Projection_2 (IT_TRANS)	Join_1 (IT_JOIN)	Star Join
Table	DIM_TRANS_MODES	CS_TRANSACTIONS	MODES and IT_TRANS	DIM_CUSTOMERACCOUNT, IT_JOIN and DIM_TRANS_MODES
Joins	DIRECT	DIRECT	Left Outer Join	Referential Join
Cardinality	NA	NA	NA	NA
Filters	NA	NA	NA	NA
Calculated Column (Expression Editor)	“AMOUNT” * “CALC_EFFECT”			
Input Parameters	NA	NA	NA	NA
Output	Transactions and Details of the Customer Account, Client, Branch, Credit Card and Loan			

VW_ALL_TRANS_V0



	Union	**Join**	**Star Join**
Table	CS_Transactions CS_Trasactions_Backend	IT_UNION_TRANS DIM_TRANS_MODES	IT_Join DIM_CUSTOMER ACC DIM_TRANS_MODES DIM_TIME_DIMENSION
Joins	Merge (IT_UNION_TRANS	Inner Joins (IT_JOIN)	JOIN
Output	Account_ID Credit Mode TXN_ID TXN_Date	Account_ID Credit Mode Amount TXN_ID TXN_Date	All output of DIM_CUSTOMERACCOUNT DIM_TRANS_MODES DIM_TIME_DIMENSION
Calculated Column	New TXN_ID	Tranns_Amount	RC: USER_TRANS_AMOUNT SYS_TRANS+AMOUNT

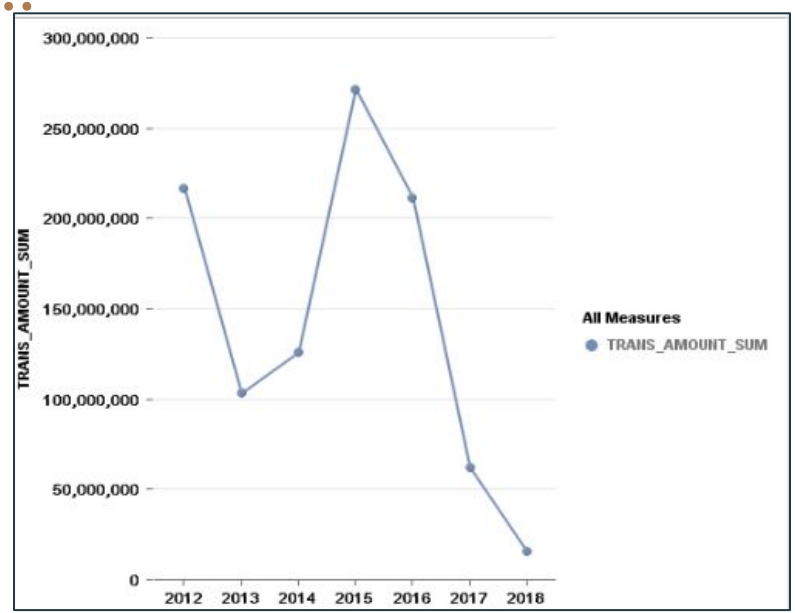
VW_ALL_TRANS_V0 ..

Primary keys: Account_ID, Mode, Date SQL

Time Range: Jan 2012- Jan 2018

System Transactions remains insignificant compared to User activated transactions

Missing data reflected “NULL” in Credit Type, Loan



Collection from another b...	14,116,608,767.5
Credit card withdrawal	-4,091,270,505.5
Credit in cash	4,844,377,545
Remittance to another bank	-13,756,659,241
Withdrawal in cash	-106,156,500

Table Selection - Summary

Table Views	View Name	Tables Used
Organisation structure & locations	DIM_BRANCH	CS_BRANCH, CS_CITY, CS_BRANCHTYPE
Customer	DIM_CUSTOMER	CS_CLIENT
	DIM_CUSTOMERACCOUNT	DIM_CUSTOMER, CS_MAPPING__TABLE, CS_ACCOUNT, CS_BRANCH, CS_CREDITCARD, CS_LOAN
STAFF RESOURCE, PAYROLL	DIM_DESIGNATIONATION	CS_DESIGNATION, CS_DESIGNATIONTEXT
	DIM_DEPATMENT	CS_DEPT, CS_DEPTEXT, CS_EMPLOYEE
	DIM_EMPLOYEE	CS_EMPLOYEE, CS_CITY
	VW_HR_EMPLOYEE	CS_EMPLOYEE, DIM_EMPLOYEE
	VW_HR_EMPSAL	CS_EMP-CTC, CS_EMP_SALCHANGE, DIM_EMPLOYEE
BANK TRANSACTIONS	VW_TRANSACTIONS	CS_TRANSACTIONS, DIM_CUSTOMERACCOUNT, CS_TRANS_MODES, CS_TRANSACTIONS_BACKEND
ACCOUNTING CUBES	DIM_GLAACCOUNT	VT_GLMASTER, CS_GLCONFIG, VT_GLCODE,
	VW_GLPOSTING	VT_CS_ACCOUNTING, DIM_GLACCOUNT

Introduction to double entry bookkeeping

Cash	
(1) $1000	(2) $1000
(3) $2000	

Equity	
	(1) $1000

Inventory	
(2) $1000	(3) $1000

Revenue	
	(3) $2000

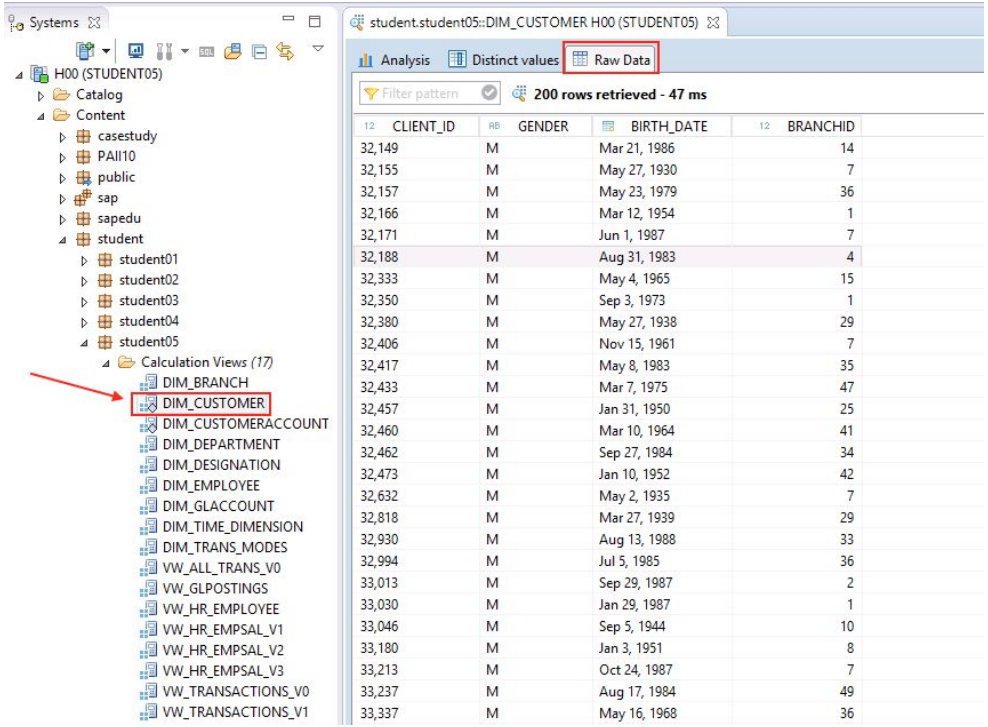
1. Shareholders put in $1000 into company
2. Company purchases $1000 worth of products
3. Company sells product for $2000

COGS	
(3) $1000	

What is a Schema?

Dimension

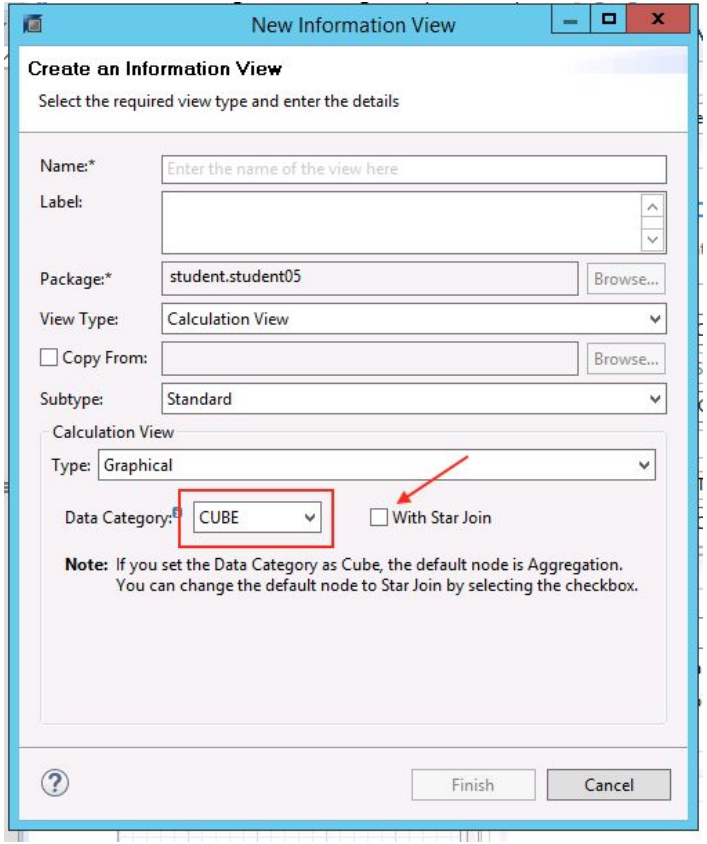
- e.g. DIM_BRANCH, DIM_CUSTOMER …
- contain only attributes
- change column names to meaningful way
- can define the table, add joins, set filters, etc.
- consumed by Cube or SQL access and re-usable



CLIENT_ID	GENDER	BIRTH_DATE	BRANCHID
32,149	M	Mar 21, 1986	14
32,155	M	May 27, 1930	7
32,157	M	May 23, 1979	36
32,166	M	Mar 12, 1954	1
32,171	M	Jun 1, 1987	7
32,188	M	Aug 31, 1983	4
32,333	M	May 4, 1965	15
32,350	M	Sep 3, 1973	1
32,380	M	May 27, 1938	29
32,406	M	Nov 15, 1961	7
32,417	M	May 8, 1983	35
32,433	M	Mar 7, 1975	47
32,457	M	Jan 31, 1950	25
32,460	M	Mar 10, 1964	41
32,462	M	Sep 27, 1984	34
32,473	M	Jan 10, 1952	42
32,632	M	May 2, 1935	7
32,818	M	Mar 27, 1939	29
32,930	M	Aug 13, 1988	33
32,994	M	Jul 5, 1985	36
33,013	M	Sep 29, 1987	2
33,030	M	Jan 29, 1987	1
33,046	M	Sep 5, 1944	10
33,180	M	Jan 3, 1951	8
33,213	M	Oct 24, 1987	7
33,237	M	Aug 17, 1984	49
33,337	M	May 16, 1968	36

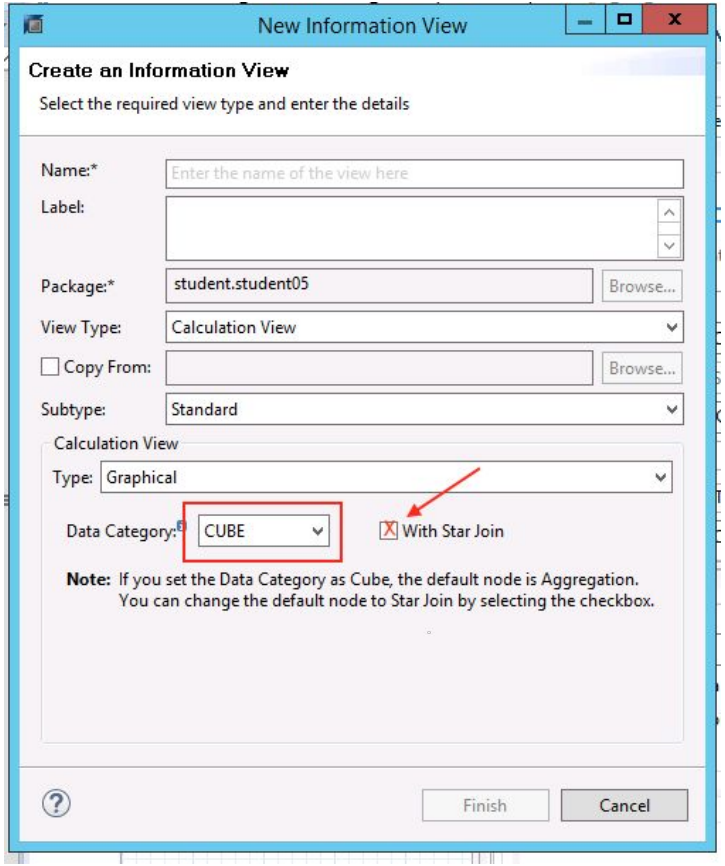
Cube

- e.g. VW_TRANSACTION_V0
- contain attributes & measures but no dimension added
- can define the table, add joins, set filters and also aggregation
- slide and dice the data possible
- consumed by reporting tools and SQL

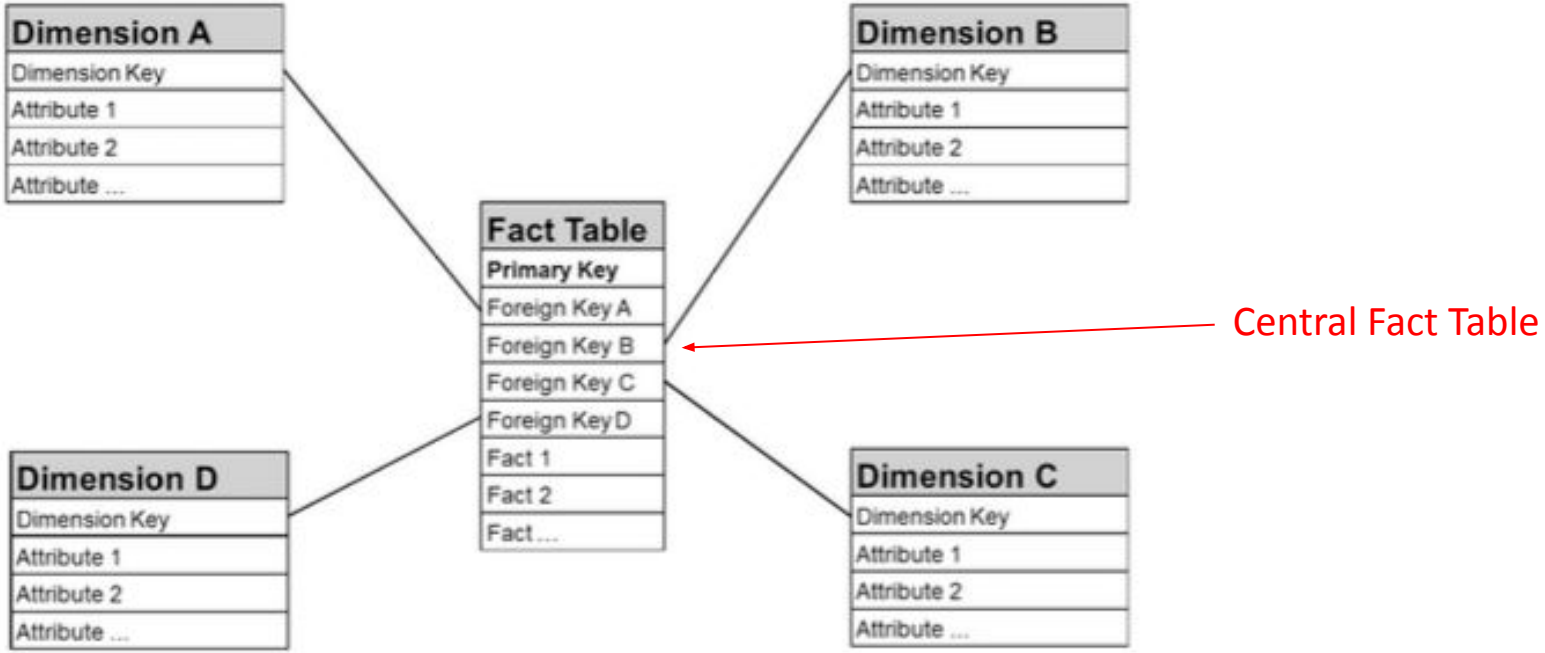


Cube with Star Join

- with Star Join
- with one or more dimension calculation views joined
- drill-down & slice and dice analysis possible
- consumed by reporting tools and SQL



Star Schema example

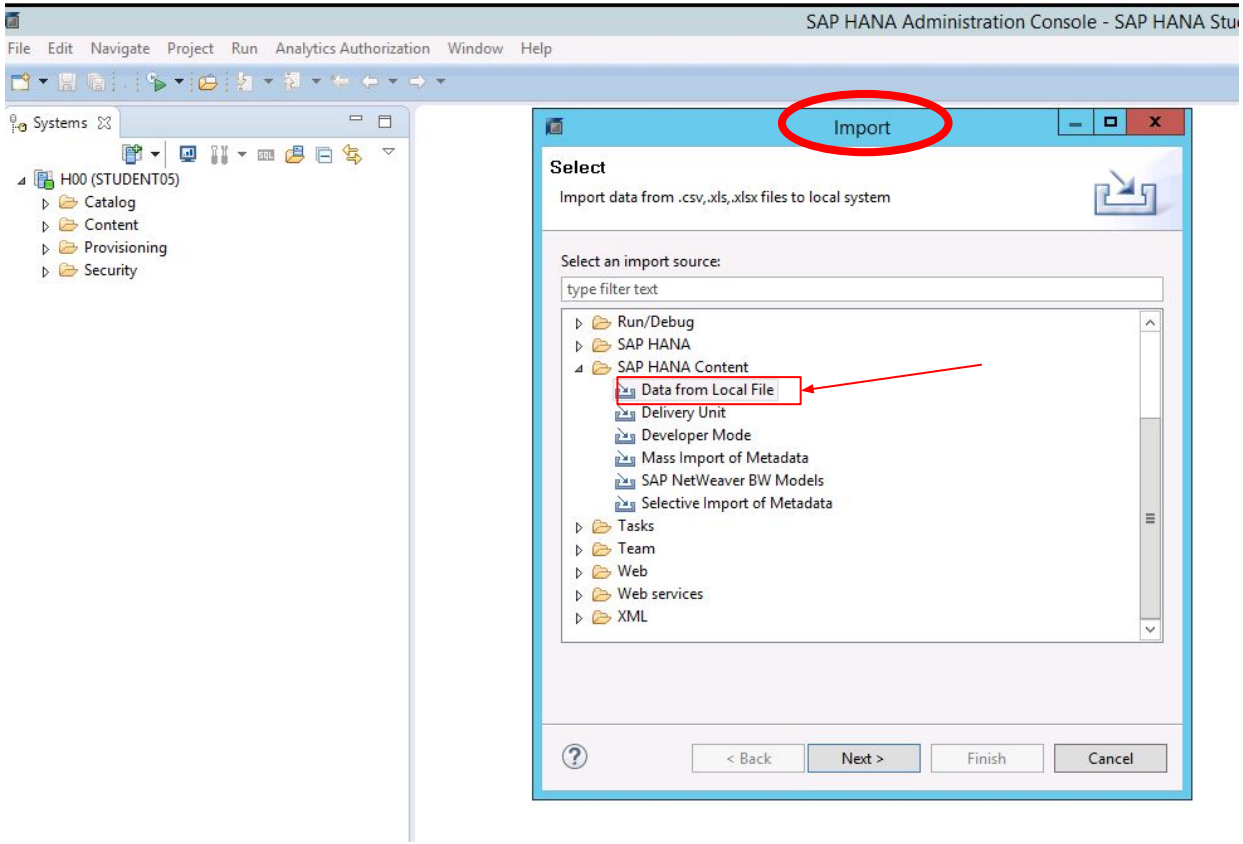


Prepare & Upload Data to SAP HANA

We are given a column table

CS_GL_CONFIG

	A	B	C	D	E	F	G
1	ACCTYPE	TRANS	DRCR	DRCR2	LINE_NUM	GL	
2	CHECKING	C1	C	D	10	2100	
3	CHECKING	C1	C	C	20	1010	
4	CHECKING	C2	C	D	10	2010	
5	CHECKING	C2	C	C	20	1010	
6	CHECKING	D1	D	D	10	1010	
7	CHECKING	D1	D	C	20	2100	
8	CHECKING	D2	D	D	10	1010	
9	CHECKING	D2	D	C	20	2110	
10	CHECKING	D3	D	D	10	1010	
11	CHECKING	D3	D	C	20	2010	
12	CHECKING	X1	D	D	10	1010	
13	CHECKING	X1	D	C	20	4020	
14	CHECKING	Y1	C	D	10	3110	
15	CHECKING	Y1	C	C	20	1010	
16	CREDITCARD	C1	C	D	10	2100	
17	CREDITCARD	C1	C	C	20	2220	
18	CREDITCARD	C2	C	D	10	2010	
19	CREDITCARD	C2	C	C	20	2220	
20	CREDITCARD	D1	D	D	10	2220	
21	CREDITCARD	D1	D	C	20	1200	
22	CREDITCARD	D2	D	D	10	2220	
23	CREDITCARD	D2	D	C	20	2110	
24	CREDITCARD	D3	D	D	10	2220	
25	CREDITCARD	D3	D	C	20	2010	
26	CREDITCARD	X1	D	D	10	2220	
27	CREDITCARD	X1	D	C	20	4030	
28	CREDITCARD	Y2	D	D	10	2220	
29	CREDITCARD	Y2	D	C	20	4120	
30	LOAN	C1	C	D	10	2100	
31	LOAN	C1	C	C	20	2210	
32	LOAN	C2	C	D	10	2010	
33	LOAN	C2	C	C	20	2210	
34	LOAN	D1	D	D	10	2210	
35	LOAN	D1	D	C	20	2110	
36	LOAN	D2	D	D	10	2210	
37	LOAN	D2	D	C	20	2110	
38	LOAN	D3	D	D	10	2210	
39	LOAN	D3	D	C	20	2010	


SAP HANA Administration Console - SAP HANA Stu
File Edit Navigate Project Run Analytics Authorization Window Help
Systems
H00 (STUDENT05)
Catalog
Content
Provisioning
Security
Import
Select
Import data from .csv,.xls,.xlsx files to local system
Select an import source:
type filter text
Run/Debug
SAP HANA
SAP HANA Content
Data from Local File
Delivery Unit
Developer Mode
Mass Import of Metadata
SAP NetWeaver BW Models
Selective Import of Metadata
Tasks
Team
Web
Web services
XML
< Back
Next >
Finish
Cancel

Import Data from Local File

Define Import Properties

Select the CSV or Excel File(Microsoft Office 97 or above) to upload data

Source File

Select File: C:\Users\train-15\Desktop\GLCONFIG.xlsx [Browse...]

File Encoding: Default

File Details

Field Delimiter: Comma (,)

Select Worksheet: Sheet1

☑ Header row exists

Header Row : 1

☐ Import all data

Start Line 2 End Line

☑ Ignore leading and trailing white-space(s) in file

Target Table

○ New Schema:

Table Name:

◉ Existing "STUDENT05".CS_GL_CONFIG [Select table]

< Back | Next > | Finish | Cancel

Import Data from Local File

Manage Table Definition and Data Mappings

Manage table definition for new tables and source target column mappings

Source File(GLCONFIG.xlsx)

Proposed Table Structure

Name
ACCTYPE
TRANS
DRCR
DRCR2
LINE_NUM
GL

Target Table(STUDENT05.CS_GL_CONFIG)

Store Type: Column store

Name	DataType	Length	S..	Key	Not...	Defa...	Mapping	Desc...
ACCOUNT_TYPE	NVARCHAR	10		☑	☑		ACCTYPE-...	
TRANSMODE	NVARCHAR	2		☑	☑		TRANS-->...	
SOURCE_DRCR	NVARCHAR	1		☑	☑		DRCR-->S...	
GLDRCR	NVARCHAR	1		☐	☐		DRCR2-->...	
LINE_NUM	TINYINT			☑	☑		LINE_NU...	
GLCODE	INTEGER			☑	☑		GL-->GLC...	

File Data : GLCONFIG.xlsx

ACCTYPE	TRANS	DRCR	DRCR2	LINE_NUM	GL
CHECKING	C1	C	D	10.0	21C
CHECKING	C1	C	C	20.0	101
CHECKING	C2	C	D	10.0	201
CHECKING	C2	C	C	20.0	101
CHECKING	D1	D	D	10.0	101
CHECKING	D1	D	C	20.0	21C

< Back

Systems

H00 (STUDENT05)
- Catalog
- Content
  - cases
  - PAII1
  - publi
  - sap
  - sape
  - stude
  - syste
- Provisior
- Security

*student.student05::VW_GL

student.student05::

- Configuration and Monitoring
- Lifecycle Management
- Backup and Recovery
- Security
- Open SQL Console
- SAP HANA Modeler
- New
- Add System with Different User...
- Add Cloud System...
- Remove — Delete
- Log Off
- Refresh — F5
- Properties — Alt+Enter

Import Data from Local File

Summary

Click Finish to confirm

Details

Source file name: C:\Users\train-15\Desktop\GLCONFIG.xlsx
Target Table Name: CS_GL_CONFIG
Target Table Schema: STUDENT05

Data from File

ACCOUNT_TYPE	TRANSMODE	SOURCE_DRCR	GLDRCR	LINE_NUM	GLCODE
CHECKING	C1	C	D	10.0	2100.0
CHECKING	C1	C	C	20.0	1010.0
CHECKING	C2	C	D	10.0	2010.0
CHECKING	C2	C	C	20.0	1010.0
CHECKING	D1	D	D	10.0	1010.0
CHECKING	D1	D	C	20.0	2100.0
CHECKING	D2	D	D	10.0	1010.0
CHECKING	D2	D	C	20.0	2110.0

Data from Existing Table

ACCOUNT_TYPE	TRANSMODE	SOURCE_DRCR	GLDRCR	LINE_NUM	GLCODE
CHECKING	C1	C	D	10	2100
CHECKING	C1	C	C	20	1010
CHECKING	C2	C	D	10	2010
CHECKING	C2	C	C	20	1010
CHECKING	D1	D	D	10	1010
CHECKING	D1	D	C	20	2100
CHECKING	D2	D	D	10	1010
CHECKING	D2	D	C	20	2110

< Back | Next > | Finish | Cancel

Data is restricted at the source ???

Data is restricted at the source ???

Data is confidential

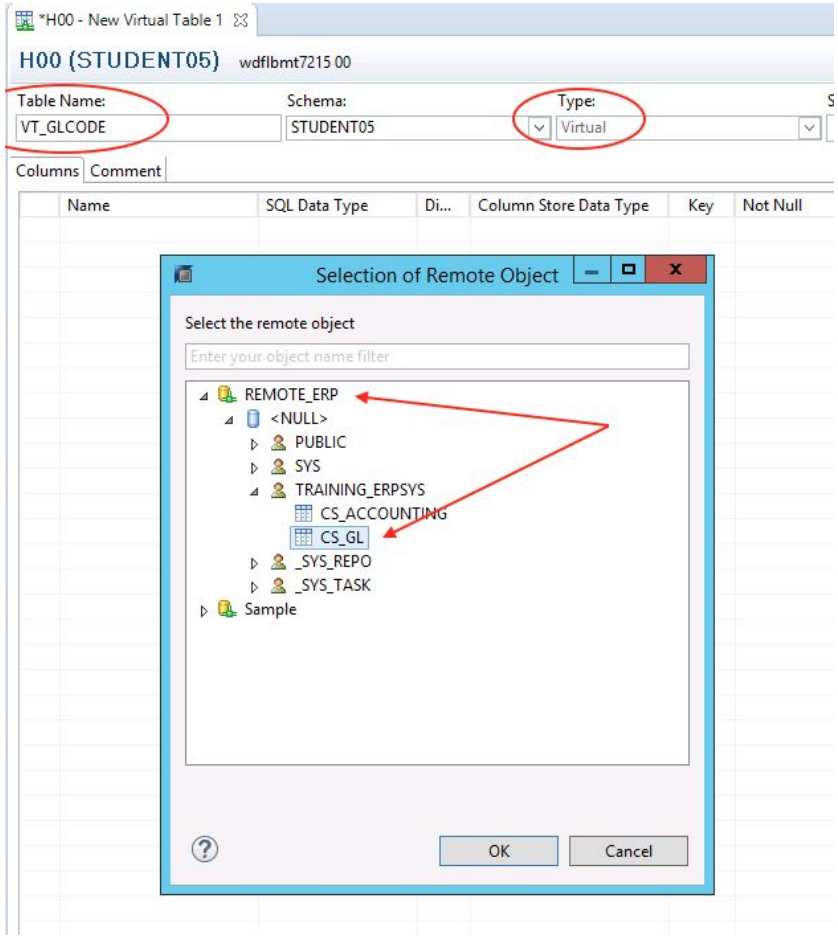
Complex calculations involved

Compliance in some countries

Solution

Remote Sources & Virtual Tables

Create Remote Table



Grant Access

*H00 - SQL Console 1

H00 (STUDENT05) wdflbmt7215 00

SQL

```
GRANT SELECT ON SCHEMA STUDENT## TO _SYS_REPO WITH GRANT OPTION;
```

Which technology the remote system is basing on

"PUBLIC"."REMOTE_SOURCES"

Raw Data | Distinct values | Analysis

Filter pattern

2 rows retrieved - 32 ms

REMOTE_SOURCE_NAME	ADAPTER_NAME	CONNECTION_INFO	LOCATION
Sample	aseodbc	<?xml version="1.0" encod...	indexserver
REMOTE_ERP	hanaodbc	<?xml version="1.0" encod...	indexserver

H00 - ADAPTERS | "PUBLIC"."ADAPTERS"

Raw Data | Distinct values | Analysis

Filter pattern | 15 rows retrieved - 47 ms | Execute | Add filter

ADAPTER_NAME	PROPERTIES	CONFIGURATION	IS_SYSTEM_ADAPTER	IS_ESS_DEFINITION_SUPPORTED
aseodbc	display_name=ASE (ODBC);description=Sybase ASE	<?xml version="1.0" en...	TRUE	FALSE
iqodbc	display_name=IQ (ODBC);description=Sybase IQ	<?xml version="1.0" en...	TRUE	FALSE
hanaodbc	display_name=HANA (ODBC);description=Hana	<?xml version="1.0" en...	TRUE	FALSE
tdodbc	display_name=TERADATA (ODBC);description=Teradata	<?xml version="1.0" en...	TRUE	FALSE
hiveodbc	display_name=HADOOP (ODBC);description=Hadoop	<?xml version="1.0" en...	TRUE	FALSE
oracle	display_name=ORACLE (GENERIC ODBC);description=Oracle	<?xml version="1.0" en...	TRUE	FALSE
mssql	display_name=MSSQL (GENERIC ODBC);description=Micro...	<?xml version="1.0" en...	TRUE	FALSE
db2	display_name=DB2 (GENERIC ODBC);description=DB2	<?xml version="1.0" en...	TRUE	FALSE
netezza	display_name=NETEZZA (GENERIC ODBC);description=Net...	<?xml version="1.0" en...	TRUE	FALSE
odbc	display_name=ODBC (GENERIC ODBC);description=ODBC	<?xml version="1.0" en...	TRUE	FALSE
maxdb	display_name=MAXDB (GENERIC ODBC);description=MaxDB	<?xml version="1.0" en...	TRUE	FALSE
hadoop	display_name=HADOOP (DESTINATION);description=HTT...	<?xml version="1.0" en...	TRUE	FALSE
sparksql	display_name=SPARK SQL (DESTINATION);description=HT...	<?xml version="1.0" en...	TRUE	FALSE
mii	display_name=MII (GENERIC ODBC);description=MII	<?xml version="1.0" en...	TRUE	FALSE
voraodbc	display_name=VORA (ODBC);description=Vora	<?xml version="1.0" en...	TRUE	FALSE

Supported Technology

H00 - ADAPTERS | "PUBLIC"."ADAPTERS"

Raw Data | Distinct values | Analysis

Filter pattern | 15 rows retrieved - 47 ms | Execute | Add filter

ADAPTER_NAME	PROPERTIES		SYSTEM_ADAPTER	IS_ESS_DEFINITION_SUPPORTED
aseodbc	display_name=ASE (ODBC);description=Sybase ASE			FALSE
iqodbc	display_name=IQ (ODBC);description=Sybase IQ	<?xml version="1.0" en...	TRUE	FALSE
hanaodbc	display_name=HANA (ODBC);description=Hana	<?xml version="1.0"		
tdodbc	display_name=TERADATA (ODBC);description=Teradata	<?xml version="1.0"		
hiveodbc	display_name=HADOOP (ODBC);description=Hadoop	<?xml version="1.0"		
oracle	display_name=ORACLE (GENERIC ODBC);description=Oracle	<?xml version="1.0"		
mssql	display_name=MSSQL (GENERIC ODBC);description=Micro...	<?xml version="1.0" en...	TRUE	FALSE
db2	display_name=DB2 (GENERIC ODBC);description=DB2	<?xml version="1.0" en...	TRUE	FALSE
netezza	display_name=NETEZZA (GENERIC ODBC);description=Net...	<?xml version="1.0" en...	TRUE	
odbc	display_name=ODBC (GENERIC ODBC);description=ODBC	<?xml version="1.0" en...	TRUE	
maxdb	display_name=MAXDB (GENERIC ODBC);description=MAXDB	<?xml version="1.0" en...	TRUE	
hadoop	display_name=HADOOP (DESTINATION);description=HTT...	<?xml version="1.0" en...	TRUE	
sparksql	display_name=SPARK SQL (DESTINATION);description=HT...		TRUE	FALSE
mii	display_name=MII (GENERIC ODBC);description=MII		TRUE	FALSE
voraodbc	display_name=VORA (ODBC);description=Vora		TRUE	FALSE

SAP ASE Sybase

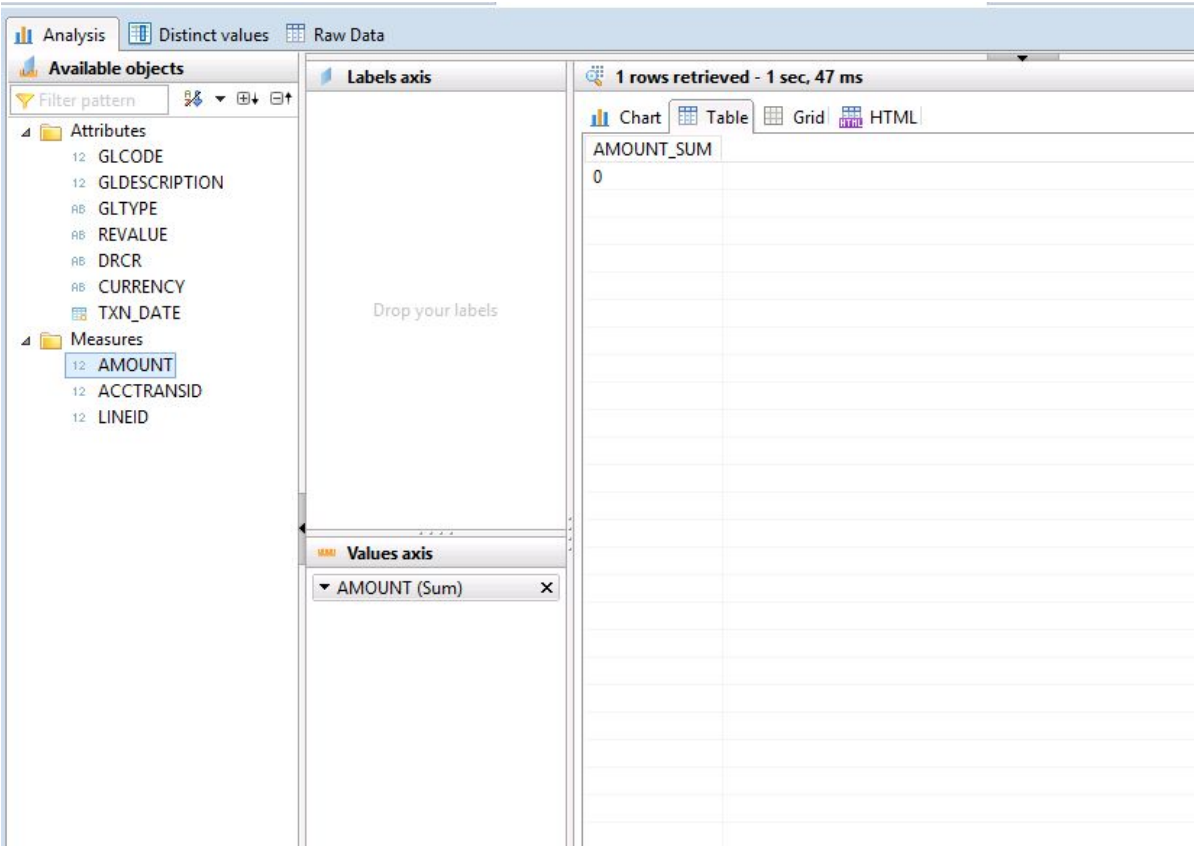
hadoop

ORACLE

Microsoft SQL Server

Analysis

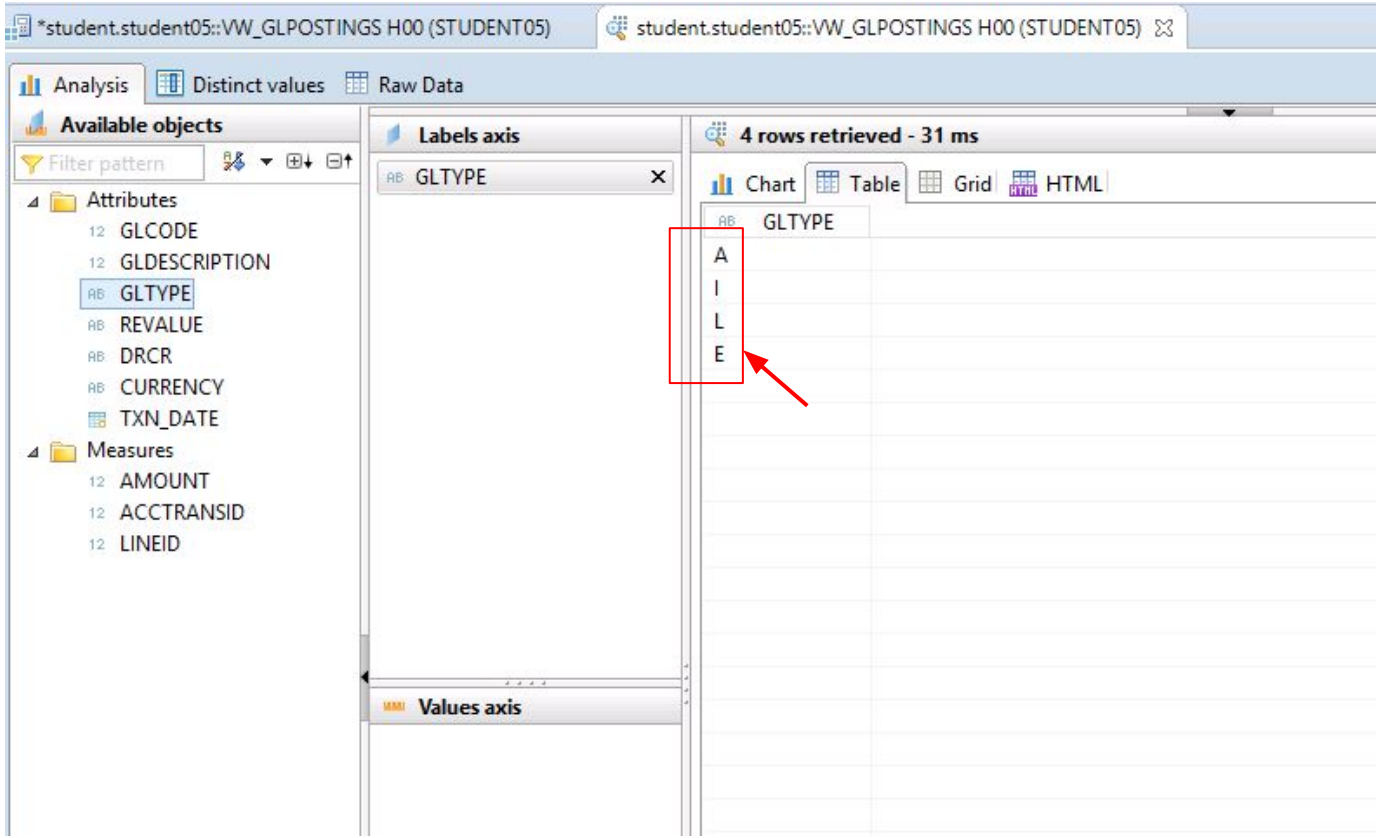
This is a counter check on proper “double entry”



Analysis

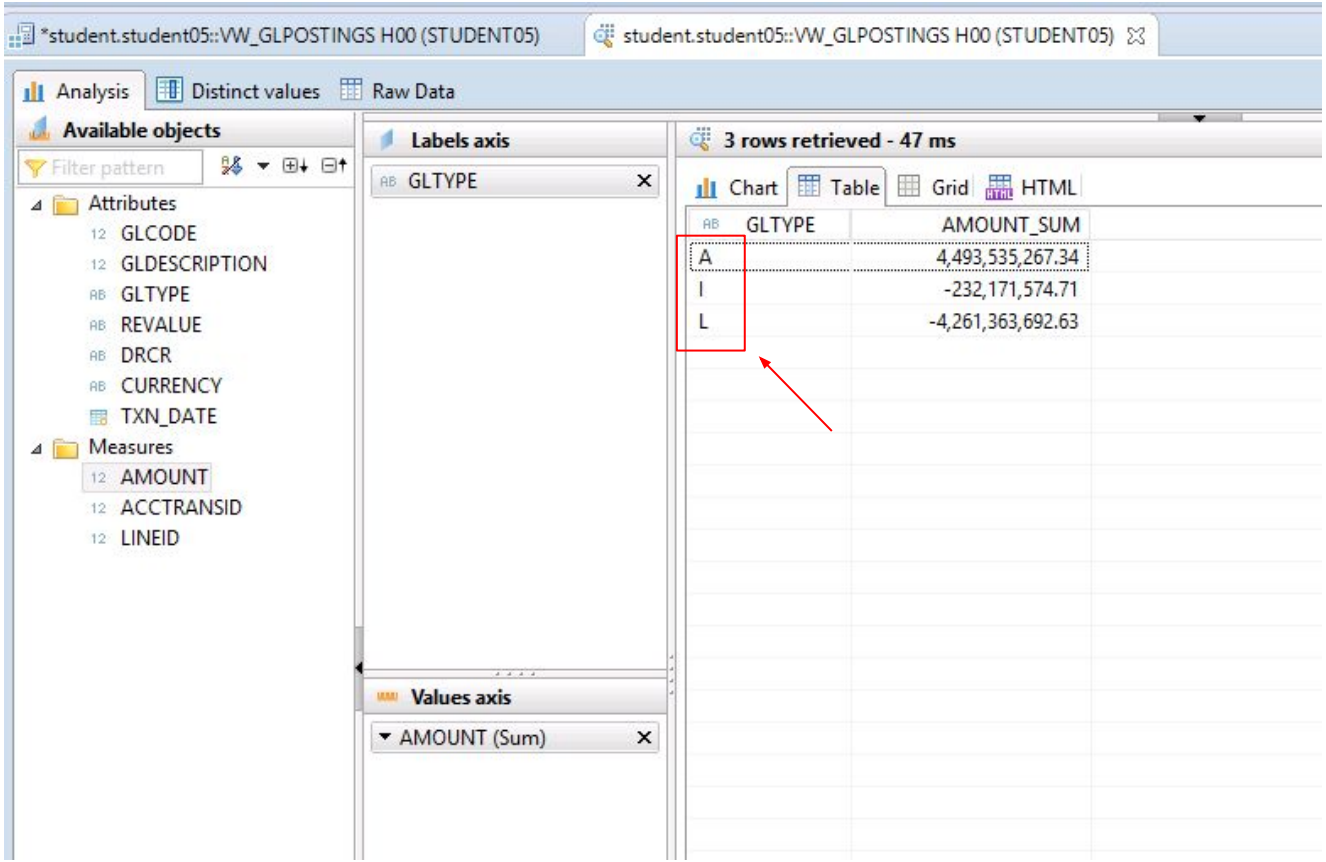
4 types of records

A, I, L, E are shown

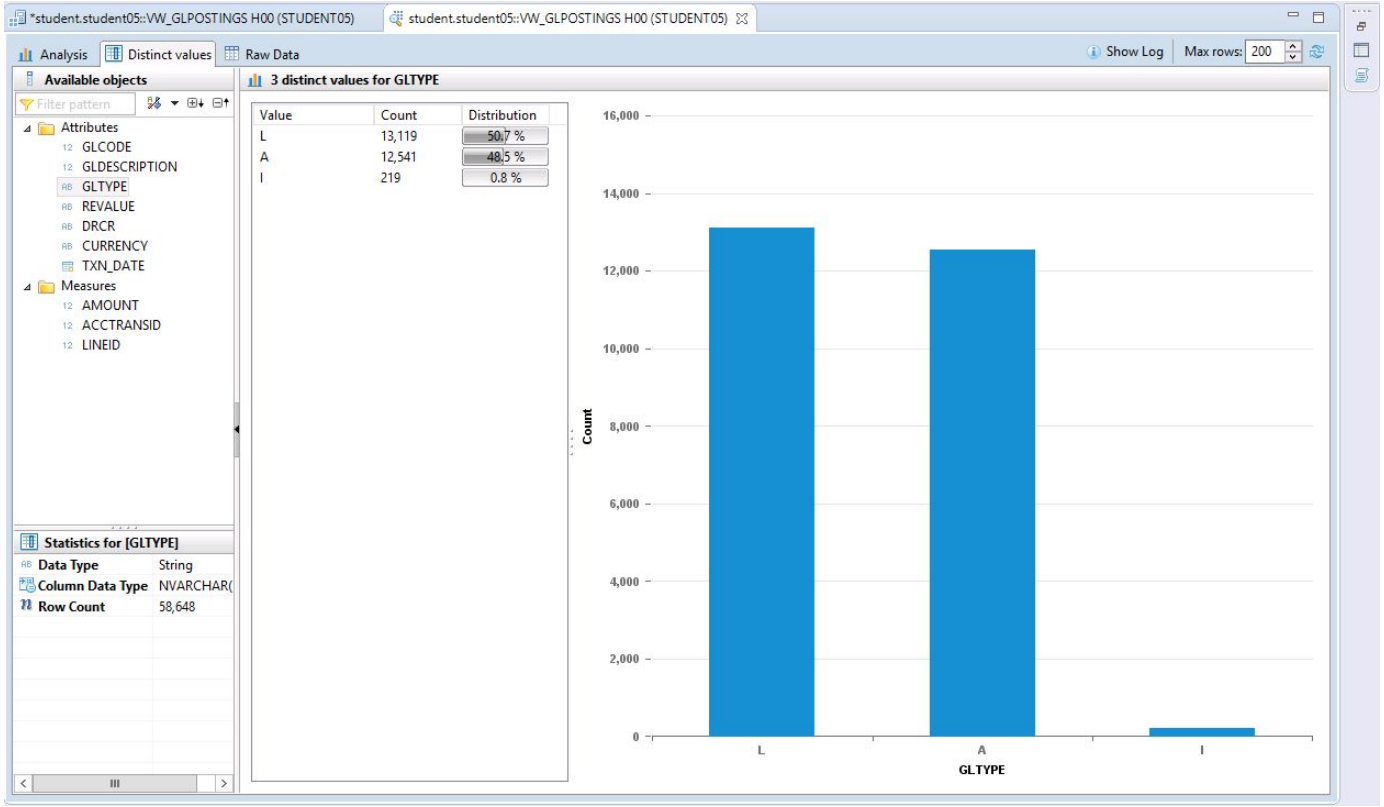


Analysis

Only 3 are shown as there is no record on E (Expenses)



Analysis

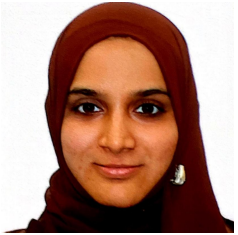

*student.student05::VW_GLPOSTINGS H00 (STUDENT05)
student.student05::VW_GLPOSTINGS H00 (STUDENT05)
Analysis
Distinct values
Raw Data
Show Log
Max rows: 200
Available objects
Filter pattern
Attributes
GLCODE
GLDESCRIPTION
GLTYPE
REVALUE
DRCR
CURRENCY
TXN_DATE
Measures
AMOUNT
ACCTRANSID
LINEID
Statistics for [GLTYPE]
Data Type String
Column Data Type NVARCHAR(
Row Count 58,648
3 distinct values for GLTYPE
Value Count Distribution
L 13,119 50.7 %
A 12,541 48.5 %
I 219 0.8 %
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Count
L
A
I
GLTYPE

END

THANK YOU

Phase 2

INTRODUCTION



Objective of Case Study

The Case Study is intended to show the team's journey using a bank setting to address a business issue. The process entails a walkthrough - a structured process of understanding the business issue, data understanding to data preparation process in a logical manner to support the ultimate findings and proposing a logical, realistic solution addressing the objectives defined.

Project Methodology- CRISP-DM

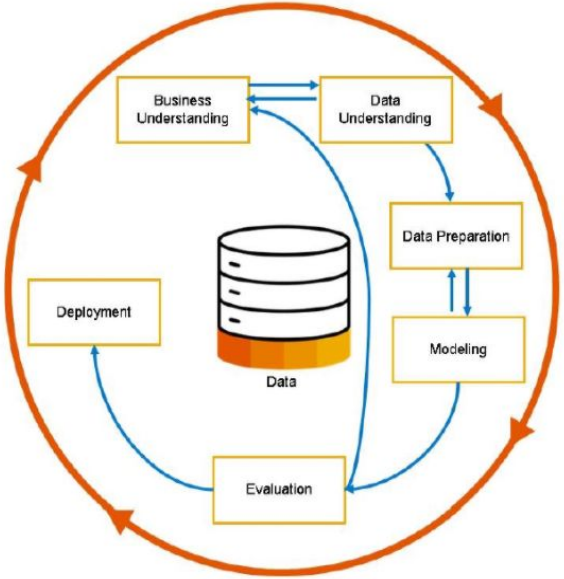
Cross-Industry Standard Process for Data Mining (CRISP-DM)

Phase1:

- Business Understanding
- Data Understanding
- Data Preparation

Phase 2:

- Modeling
- Evaluation
- Deployment



extract: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

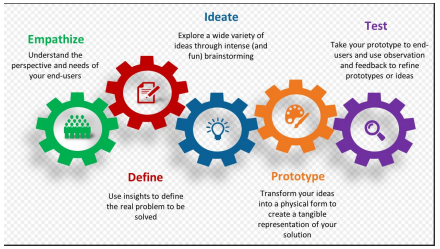
CRISP-DM Project Methodology

35 days

TIME						
Project Prep & readiness	5 days					
Business Understanding	6 days					
Data Understanding		11 days				
Modeling			6 days			
Evaluation					2 days	
Deployment						5 days

Empathize

DESIGN THINKING METHODOLOGY

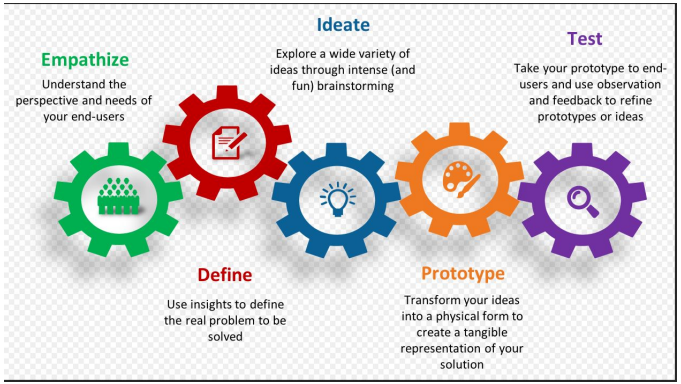


Objectives

- Understand Requirement and Stakeholder
  - From data attribute perspective, each department needs to know their department's transaction data.
- Understand Business Challenges
  - The Bank would like to expand their business in new Credit Card, but it does not know which group can be targeted.
- Understand Existing Business Processes
  - At the end of month or year, the Bank is busy on reporting by manual work. The manual work with huge data leads to loss of data accuracy and increasement on human dependency.

Empathize

Define Problem and Solution

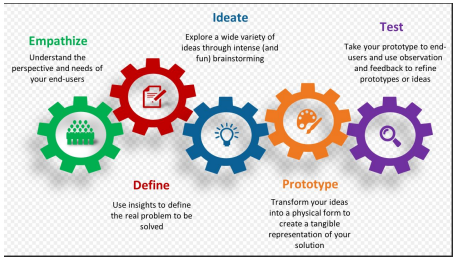


Core Problem Identification

- Identify Requirements and Stakeholders
  - Need dimensions and measures for different scenarios.
  - Different roles can have different rights to the content.
- Identify Problems with Business
  - Lack of Enterprise Database to handle multiple operations and business in globe.
  - The Bank would like to expand their business in new Credit Card, but does not know which groups should be targeted.
- Problem with Procedure
  - Leverage SAP BI scheduling and publication function to distribute analytics content to right stakeholders automatically.

Ideate

Ideation Stage

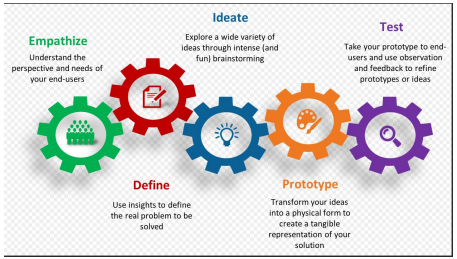


After understanding the challenges and problems identification, we shall come up with some ideas. Shall Recap what data we have had in hand and create a draft landscape.

Recap Data and Main HANA Views:

1. Branch information of the Bank: DIM_BRANCH
2. Customer information: DIM_CUSTOMER and DIM_CUSTOMERACCOUNT
3. Department information: DIM_DEPARTMENT
4. Employee information: DIM_EMPLOYEE
5. Designation information: DIM_DESIGNATION
6. GL Account information: DIM_GLACCOUNT
7. GL Posting information: DIM_GLPOSTINGS
8. Time information: DIM_TIME_DIMENSION
9. Transaction mode information: DIM_TRANS_MODES
10. Calculation View - Transaction view (plus internal): VW_ALL_TRANS
11. Calculation View – GL Postings view: VW_GLPOSTINGS
12. Calculation View – Employee view: VW_HR_EMPLOYEE
13. Calculation View – Salary view: VW_HR_EMPSAL
14. Calculation View – Transaction view (customer only): VW_TRANSACTIONS

Ideate



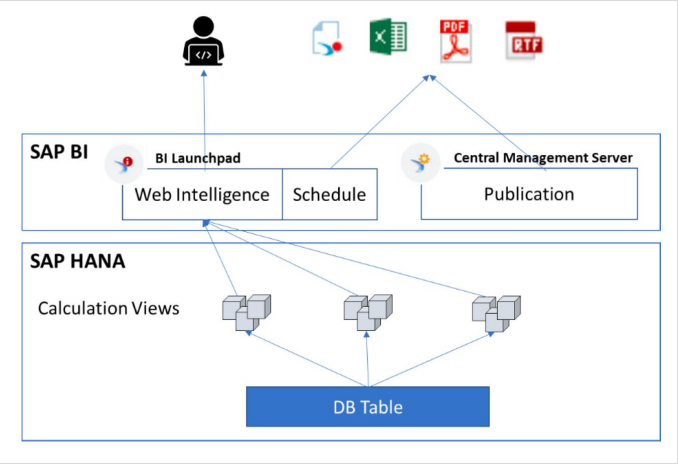
Ideation Stage

After understanding the challenges and problems identification, we shall come up with some ideas. Shall Recap what data we have had in hand and create a draft landscape.

What We Can Answer from the Views:

1. Transaction amount of each branch.
2. Customer balance variation.
3. Diversity analysis of your employees.
4. Salary analysis.
5. GL balance analysis.

What the Draft Landscape Can Be:



Role Play with 4 Personas

Name: **U-Sing Bolt**

I love sports so as to stay energetic all the day !!

Background:

Working in sales dept for 20 years

Job Title / Role:

Sales Manager

JOB RESPONSIBILITIES

Main tasks and frequency:

- Track leads
- Maintain customer relationship
- Look for new opportunities
- Initiating campaigns

MAIN GOALS

- Build strong relationship with existing customers
- Sell new products to existing customers
- Find new opportunities & new customers
- Meeting sales target

PAIN POINTS

- Involved a lot of manual work and duplications
- No intelligence on potential customer exploration
- Huge amount of data but in different formats flows in from different countries

NEEDS

- Get up-to-date info to visualize and analyze data from different sources
- To gain insights on target markets & potential customers
- Info providing potential customers & products for cross-selling
- Digital customer list for maintaining customer relations

STAKEHOLDERS

Dept heads, Customers, IT colleagues

COMPETENCIES

Casual User		Power User
Proactive		Reactive
Work in team		Lone Fighter
Global focus		Local focus
Innovative		Conservative

Name: Bo Joe

I love data!

Background:
Classics, Literature, Philosophy

Job Title / Role:
Analytical Manager

JOB RESPONSIBILITIES

Main tasks and frequency:

- Team Lead
- Strategy development
- Implementing, streamlining & execution workload allocation for data.

MAIN GOALS

- Search for latest trend and technology.
- Continuous improvement to data analytics

PAIN POINTS

- Stakeholder management to ensure their goals are align.
- Time & cost required for data cleansing.
- No headcount for new hiring.

NEEDS

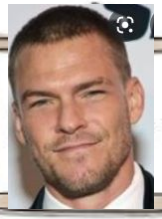
- Mission & Vision Statement must be clearly defined.
- Deliverables & Expectation must be well articulated.
- Investment in data storage infrastructure without having to rely on cloud facilities.
- Additional headcount.

STAKEHOLDERS

- Head of sales, HR Business Partner, Credit card product manager

COMPETENCIES

Casual User	Power User
Proactive	Reactive
Work in team	Lone Fighter
Global focus	Local focus
Innovative	Conservative



Name: Jack Sparrow

I haven't reported my missing credit card to the police because whoever stole it is spending less than my wife.

Background:

Promoted from sales personnel

Job Title / Role:

Credit Card Product Manager

JOB RESPONSIBILITIES

Main tasks and frequency:

Research, analyze customer and industry trends. Align with corporate strategy
Tracking, monitoring and analysis of Credit Card products data
New product or promotion launch design and development
Exploring potential innovative business opportunities
Studying of market trends, user needs, competitors' products

MAIN GOALS

Creating effective credit card products and new launches.
Working with sales team to generate new leads and customers
Ensuring current products are profitable and to decide when to axe old products

PAIN POINTS

Large amount of data to be processed and analysed.
Difficulty in getting quality data
Multiple sources of data
Hard to get accurate results from the products without extensive time spent

NEEDS

Powerful analysis tool to help sieve through numerous data and churn out highly specific results catered to his needs
Accurate ways to measure and predict outcomes for certain products before launch, so that he can make informed decisions and changes. This would lead to better numbers for the company
Cohesion between various departments to successfully launch new products. As well as getting previous data from sales team so as to analyse the data accurately.

STAKEHOLDERS

Sales Director, Marketing Director, Customers

COMPETENCIES

Casual User		Power User
Proactive		Reactive
Work in team		Lone Fighter
Global focus		Local focus
Innovative		Conservative

Name: Tina Turner

Salaries are now on my finger tips

Background:

HR Management

Job Title / Role:

HR Manager

JOB RESPONSIBILITIES

Main tasks and frequency:

-To maintain a modern and efficient HR IT System
-Payroll processing-monthly pay & year end bonus and increment
-HR reports-Monthly/Annually/Ad Hoc
-To keep track of employees
-To address staff problems quickly and efficiently

MAIN GOALS

-Ensure company complies with laws and standards for employees
-Managing benefits and payroll without delays and errors
-Fair Salary proposal based on analysis of performance
-Fast and effective reports for pay forecasting

PAIN POINTS

-Training new recruits
-Uses outdated HR Software
-Not having enough insights to identify the possible high turnover period

NEEDS

-Identify and Analyse the areas that Employees need training for
-Data Analysis of employee data
-Scalable HR Software as the business grows

STAKEHOLDERS

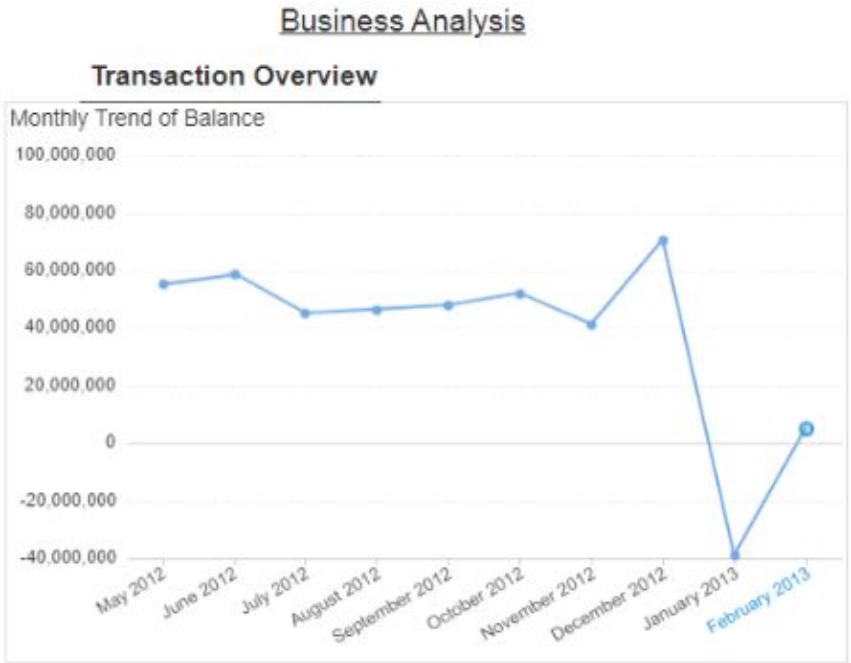
-Department Head
-Regional HR Director

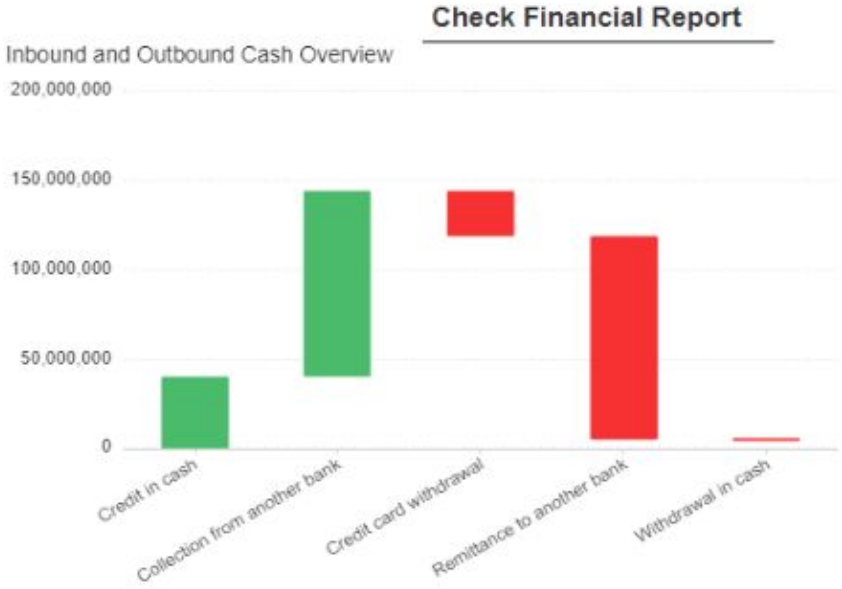
COMPETENCIES

Casual User	Power User
Proactive	Reactive
Work in team	Lone Fighter
Global focus	Local focus
Innovative	Conservative

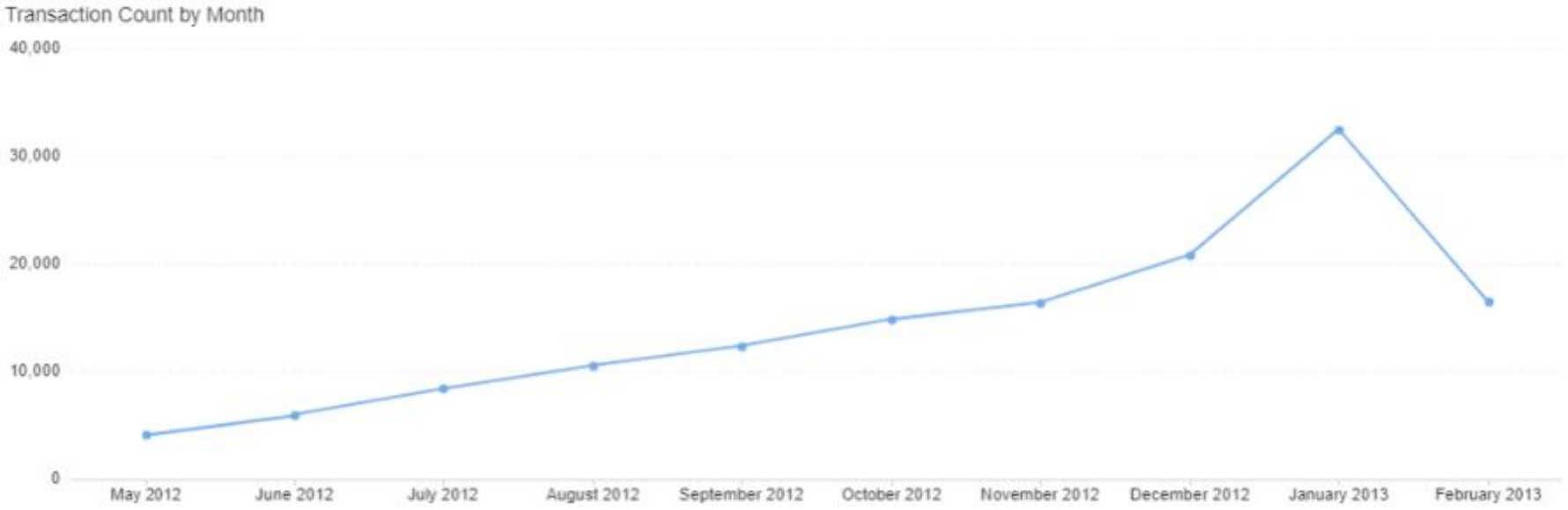
Use of BI Launch Pad for Business Understanding

Understanding our business and locating valuable customer.

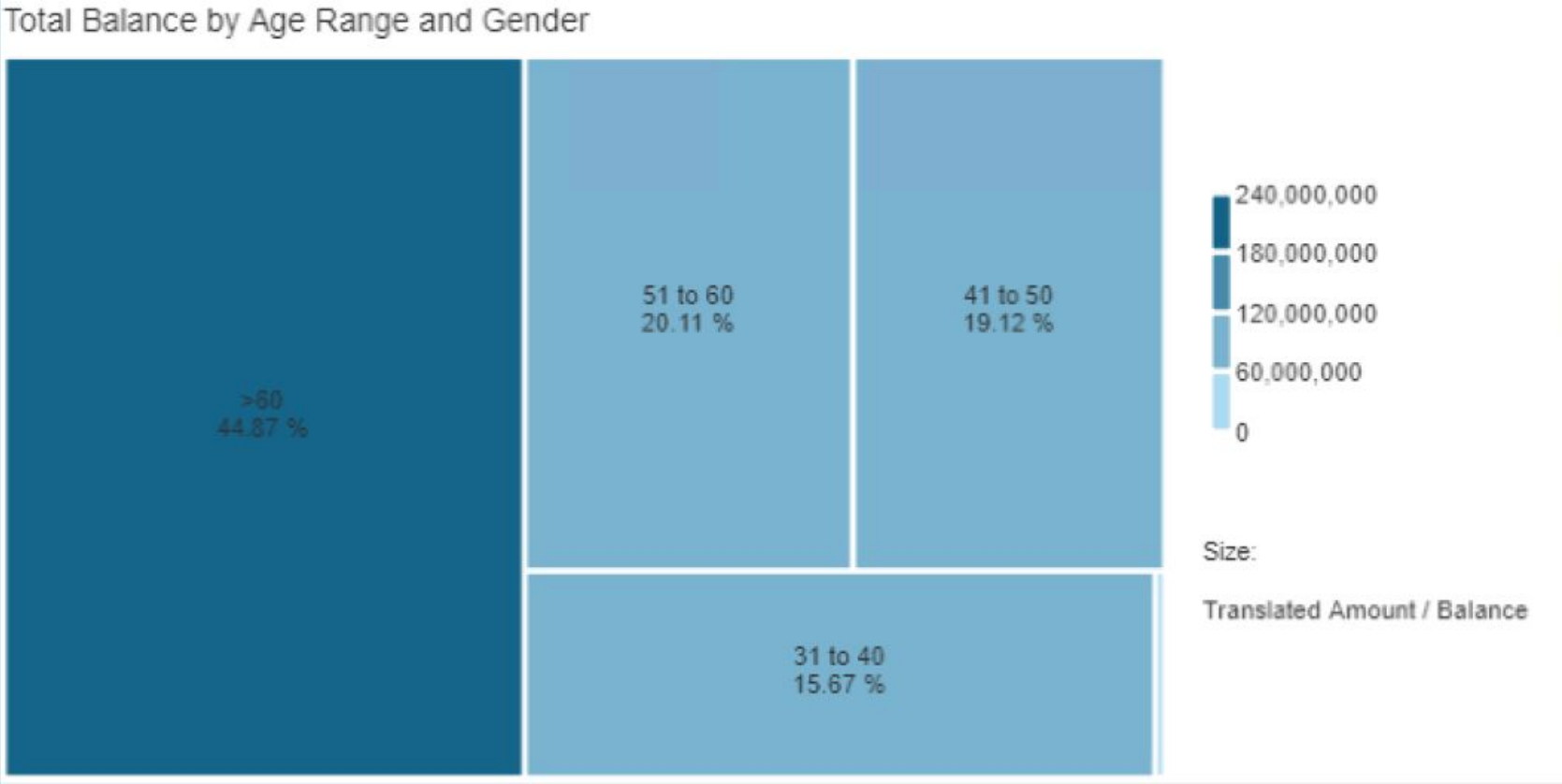




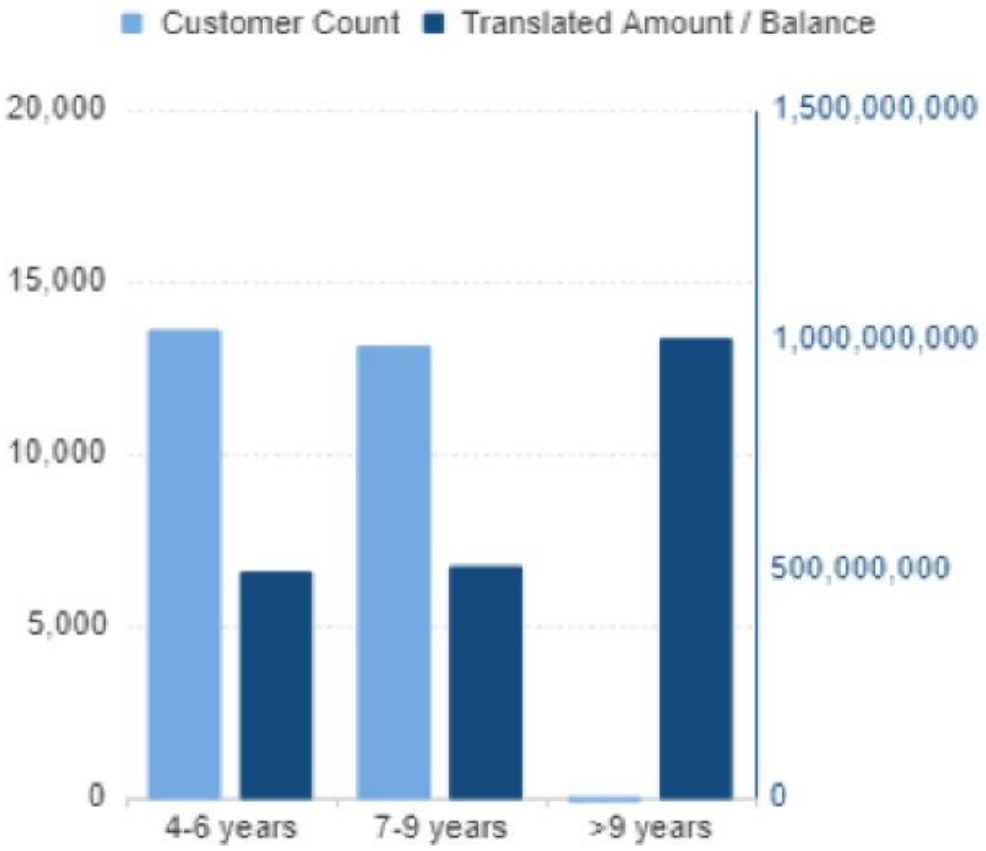
Customer Overview



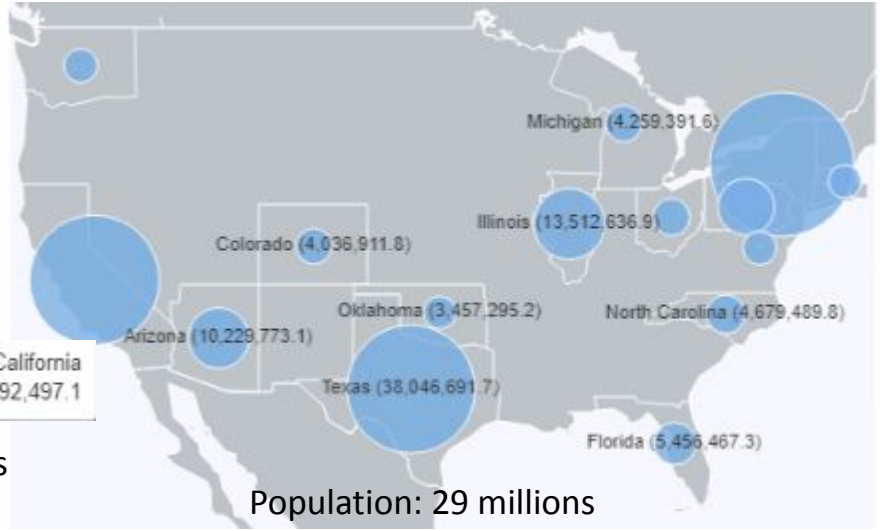
Most Valuable Customer



Retention Year & Total Balances



Location Overview



REGION: New York
Translated Amount / Balance: 46,809,904.4

Population: 8 millions

Translated Amount / Balance

REGION: California
Translated Amount / Balance: 38,992,497.1

Population: 40 millions

Population: 29 millions

Financial Reporting from 1st May 2012 thru 30th April 2013

Assets

CURRENCY	GL Code	GL Description	Amount
USD	2,010	Cash in Hand	291,059,568
USD	2,210	Loan Customer Control Account	-21,674,326.76
USD	2,220	Credit Card Customer Control Account	-35,226,939.3
		Sum:	**234,158,301.94**

Incomes

CURRENCY	GL Code	GL Description	Amount
USD	4,010	Banking Charges Income - Advances Account	-1,955,552.14
USD	4,020	Banking Charges Income - Management Fees	-5,016,645.81
USD	4,030	Banking Charges Income - Credit Card Fees	-4,413,508
		Sum:	**-11,385,705.95**

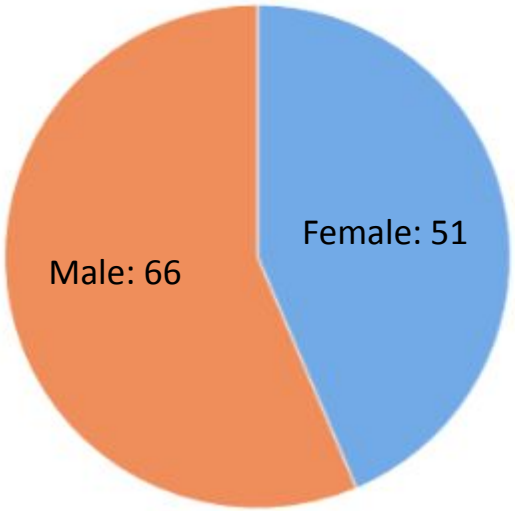
Liabilities

CURRENCY	GL Code	GL Description	Amount
USD	1,010	Checking Customer Control Account	-125,371,902.69
USD	1,200	Credit Card Affiliate Vendor Payment	-187,579,360.4
USD	2,100	Interbank / External Bank Obligations	374,856,306
USD	2,110	Interbank / External Other Withdrawal Obligations	-284,677,638.9
		Sum:	**-222,772,595.99**

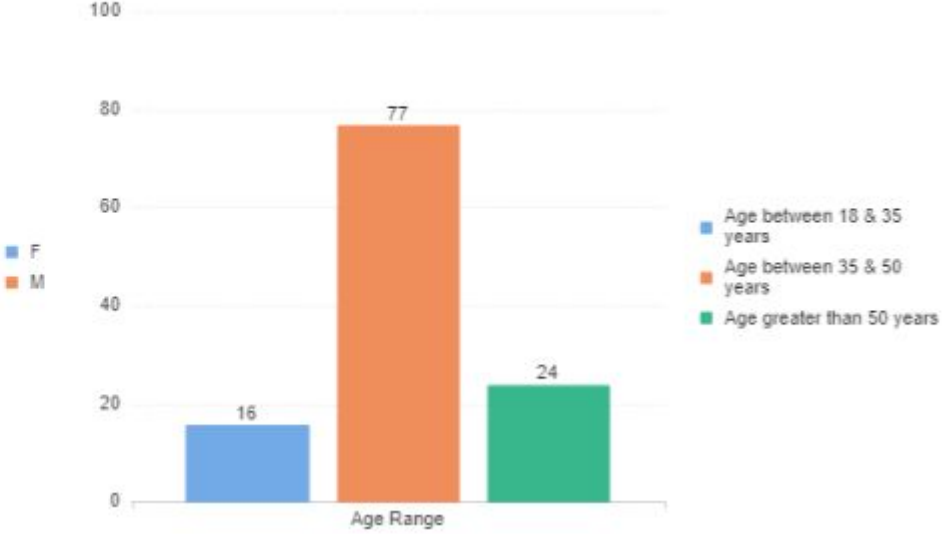
Understanding our Employee

Diversity Analysis

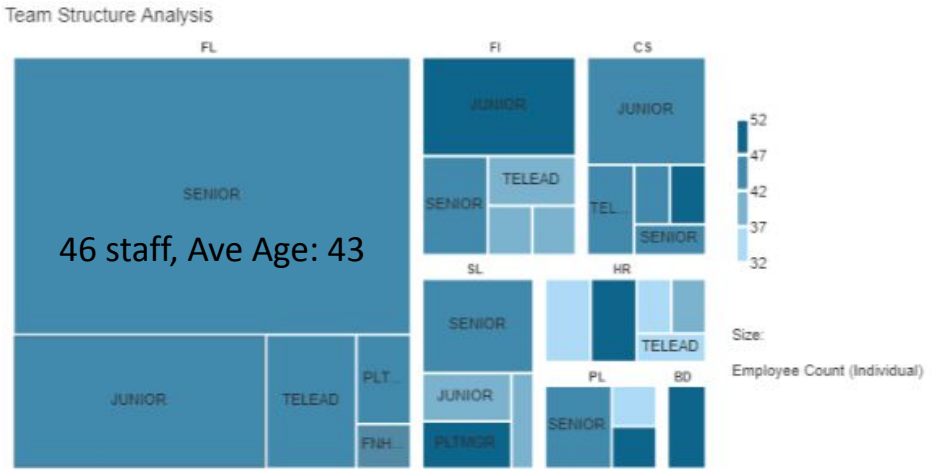
Diversity Analysis by Gender



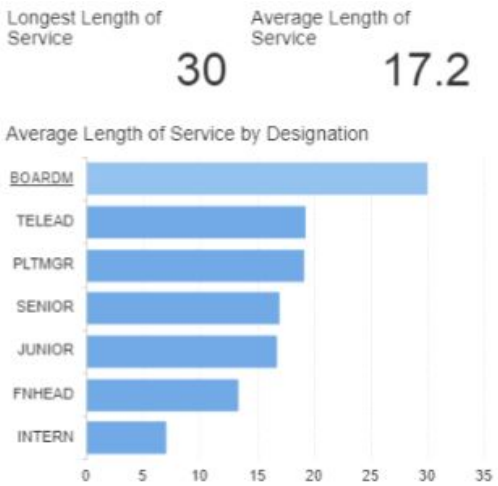
Diversity Analysis by Age



Understanding the Team Dynamics



14 staff, Ave Age: 44



11 colleagues are celebrating birthday this month

Department	Name	Gender	Date of Birth	Age	Length of Service
FL	AK	M	19790428	42	16
FL	AP	M	19870419	34	17
CS	AT	F	19800408	42	18
FL	BC	M	19960407	26	7
SL	BM	F	19730406	49	16
FL	BR	M	19850429	36	13
SL	BX	M	19800429	41	17
FL	CH	F	19820408	40	21
PL	CW	F	19760407	46	22
FL	FL	F	19640430	57	23
FL	FL	F	19860403	36	14

Yearly Salary & Wage Report 2018

Growth Rate

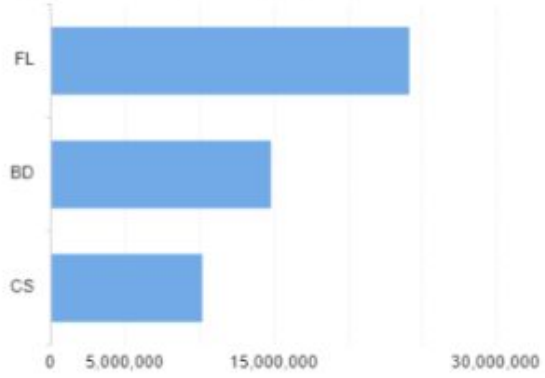
8.74%

Total Salary and Wage

Growth Rate Average

8.74%

Average Salary and Wage"

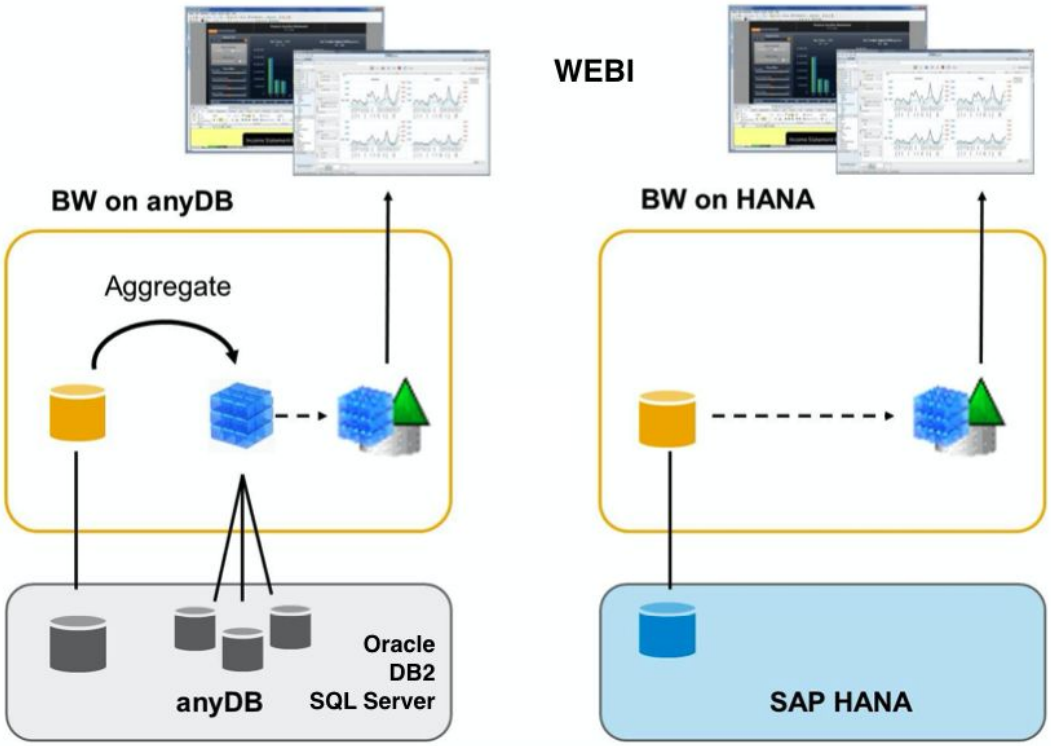


Salary and Wage Details

HIER_EMPORG	Department C	Department C	Gender	Growth Rate of Average Sa	Current Salary	Previous Year Salary	Salary 2 Years Back
Board A	BD	BD	M	8.71%	9,938,400	9,142,348	8,495,245
CA	HR	HR	F	8.59%	8,976,400	8,266,210	7,677,189
Board B	BD	BD	F	9.61%	4,861,500	4,435,169	4,128,777
AV	PL	PL	M	10.20%	981,600	890,779	827,177
AY	FI	FI	F	9.37%	1,603,900	1,466,496	1,367,617
CE	SL	SL	F	9.23%	1,376,100	1,259,803	1,169,412

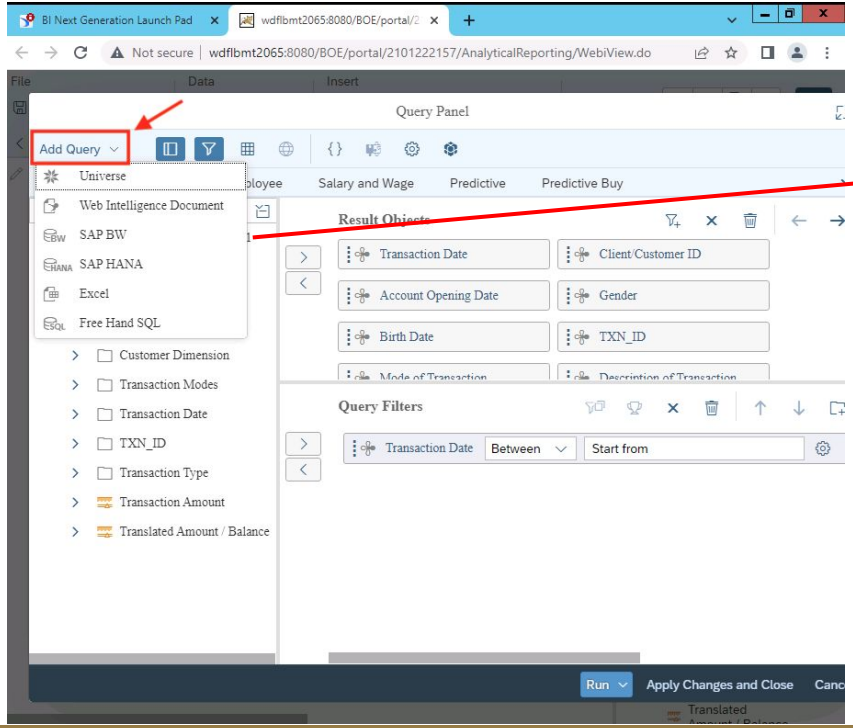
WEBI Data Blending & Document Sharing

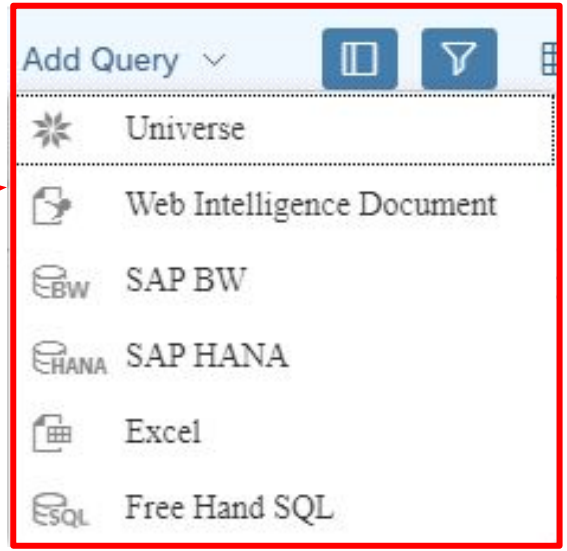
Types of Data Sources Supported

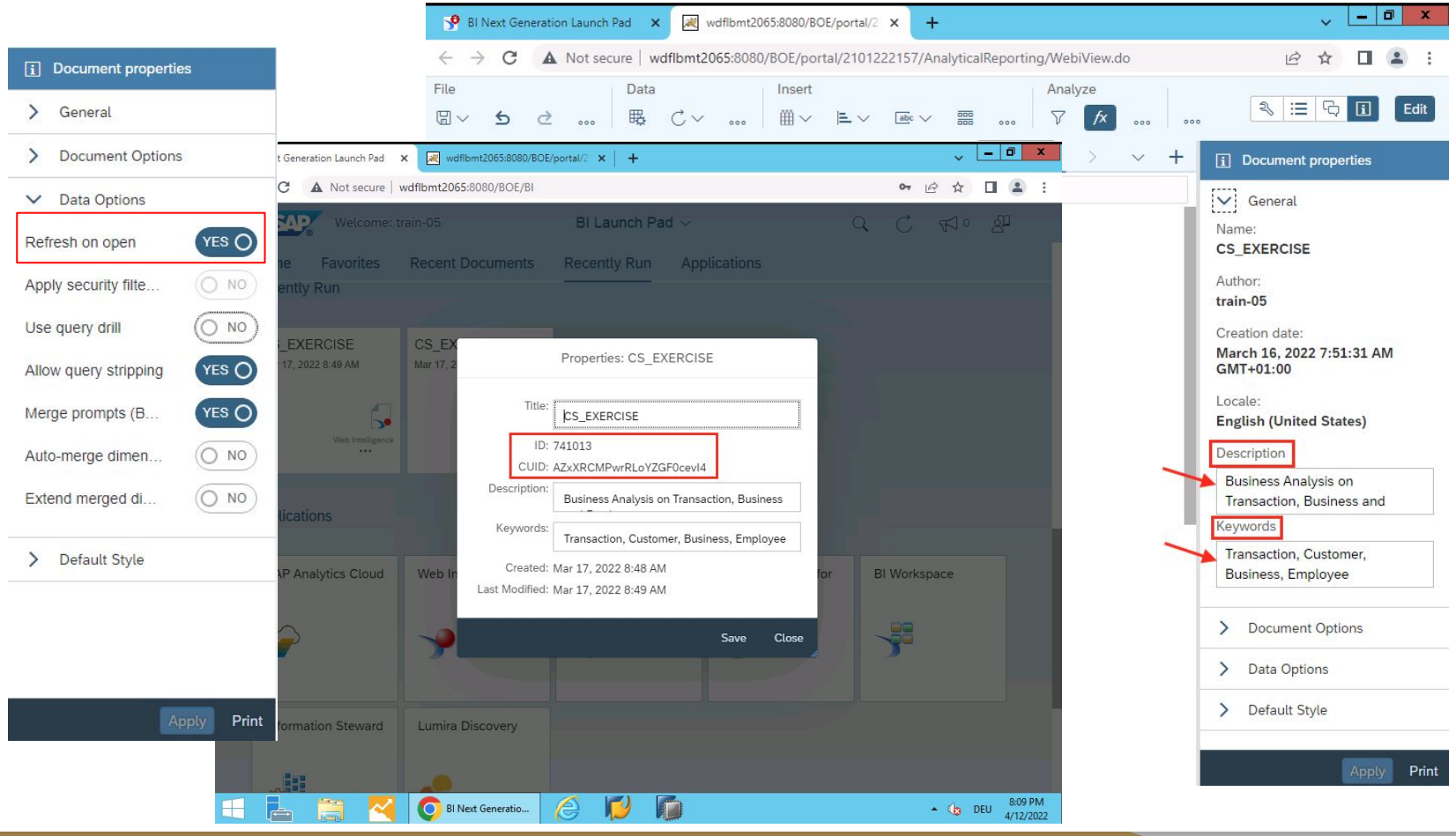


- SAP HANA
- SAP BW
- Oracle
- DB2
- SQL Server

Access to Excel and Reporting






Document properties
General
Document Options
Data Options
Refresh on open
YES
Apply security filte...
NO
Use query drill
NO
Allow query stripping
YES
Merge prompts (B...
YES
Auto-merge dimen...
NO
Extend merged di...
NO
Default Style
Apply
Print
BI Next Generation Launch Pad
wdflbmt2065:8080/BOE/portal/2
Not secure
wdflbmt2065:8080/BOE/portal/2101222157/AnalyticalReporting/WebiView.do
File
Data
Insert
Analyze
Edit
Not secure
wdflbmt2065:8080/BOE/BI
Welcome: train-05
BI Launch Pad
Favorites
Recent Documents
Recently Run
Applications
Properties: CS_EXERCISE
Title:
CS_EXERCISE
ID: 741013
CUID: AZxXRCMPwrRLoYZGF0cevI4
Description:
Business Analysis on Transaction, Business
Keywords:
Transaction, Customer, Business, Employee
Created: Mar 17, 2022 8:48 AM
Last Modified: Mar 17, 2022 8:49 AM
Save
Close
BI Workspace
Lumira Discovery
Document properties
General
Name:
CS_EXERCISE
Author:
train-05
Creation date:
March 16, 2022 7:51:31 AM GMT+01:00
Locale:
English (United States)
Description
Business Analysis on Transaction, Business and
Keywords
Transaction, Customer, Business, Employee
Document Options
Data Options
Default Style
Apply
Print

Export WEBI Document

Export WEBI Document

Financial Report

Reporting From: 2012- 05-01 To: 2013-04-30 Currency: USD

Assets

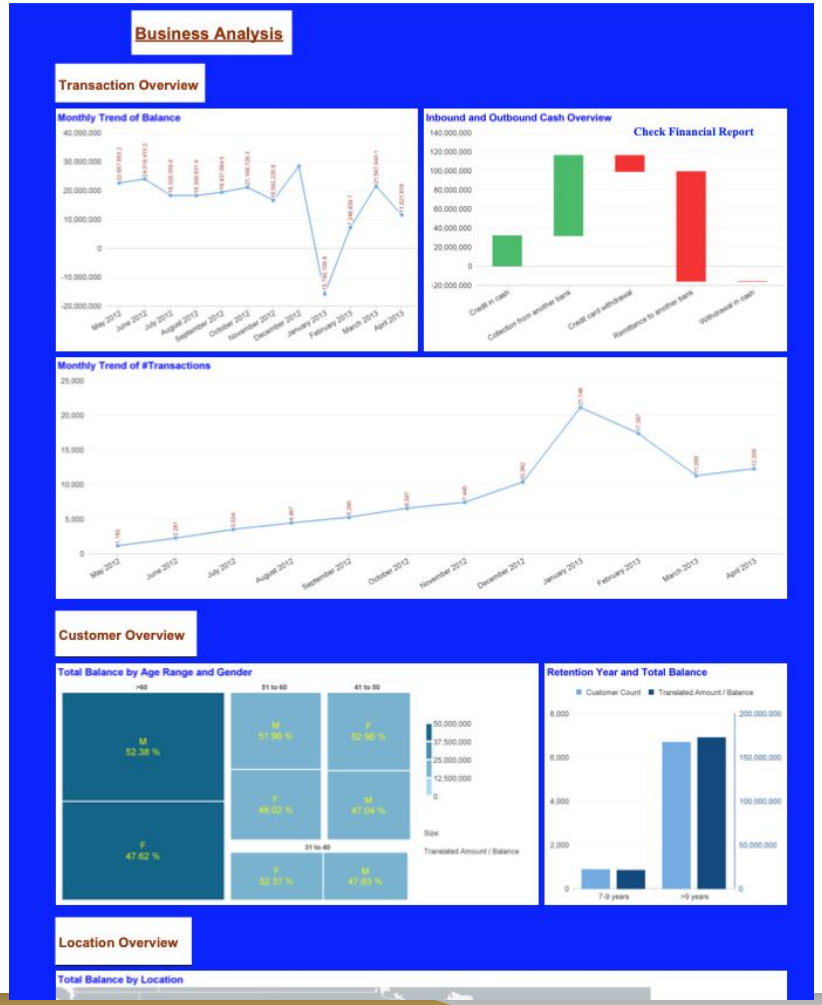
GL Code	GL Description	Amount
2010	Cash in Hand	291,059,568
2210	Loan Customer Control Account	-21,674,326.76
2220	Credit Card Customer Control Account	-35,226,939.3
	Sum:	**234,158,301.94**

Incomes

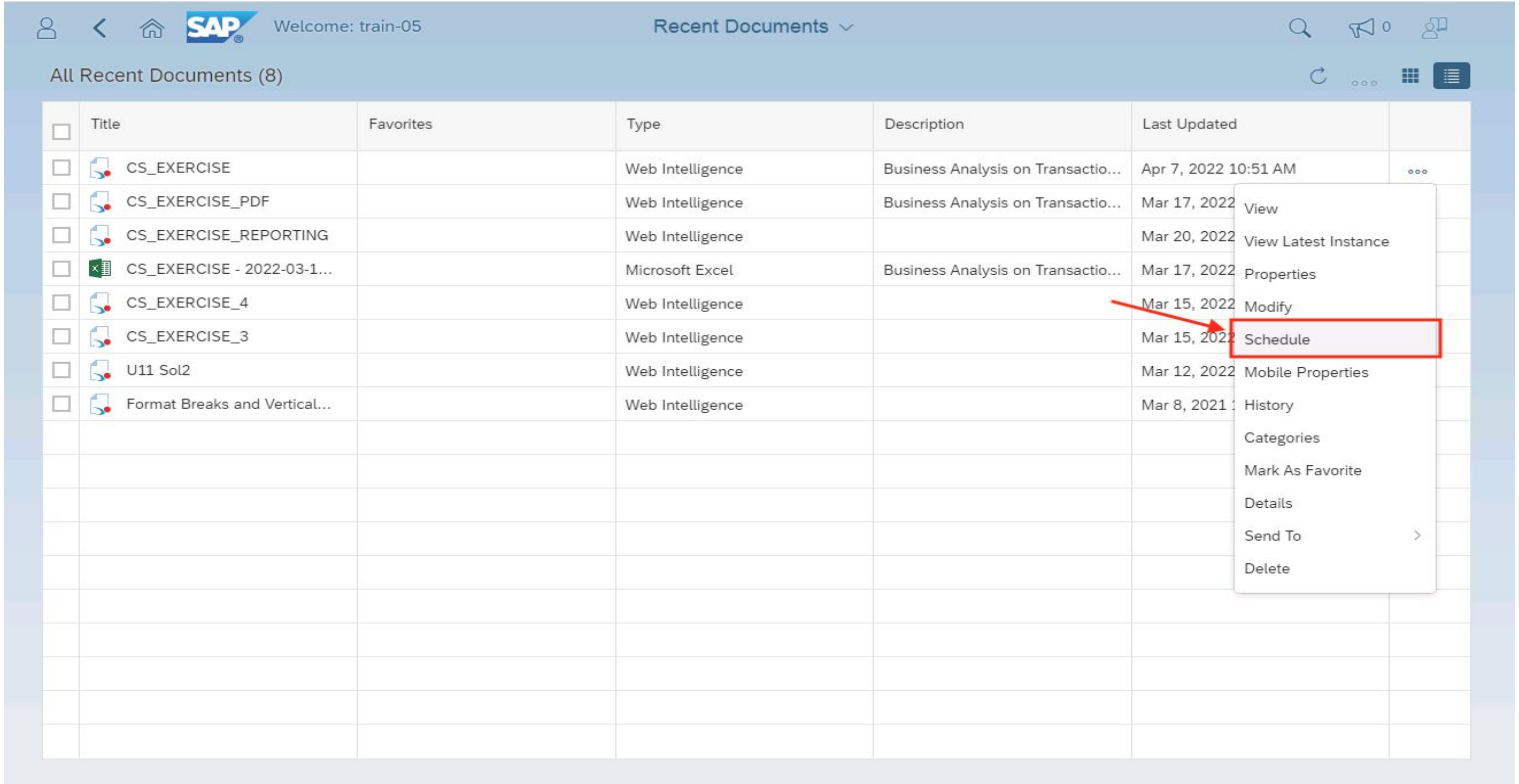
GL Code	GL Description	Amount
4010	Banking Charges Income - Advances Account	-1,955,552.14
4020	Banking Charges Income - Management Fees	-5,016,645.81
4030	Banking Charges Income - Credit Card Fees	-4,413,508
	Sum:	**-11,385,705.95**

Liabilities

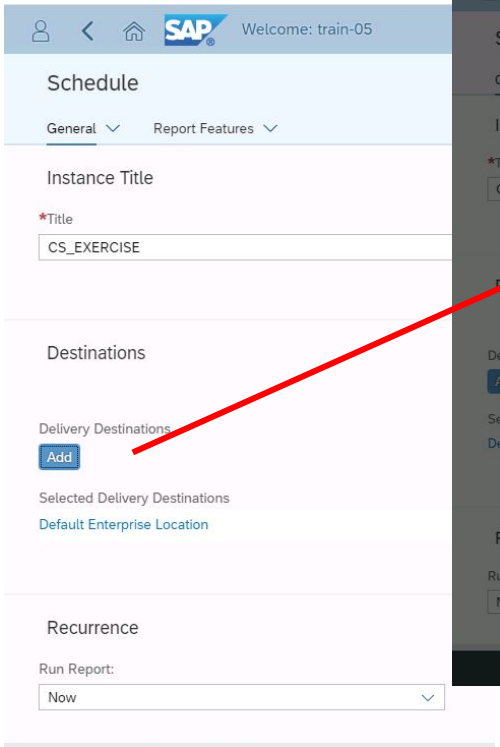
GL Code	GL Description	Amount
1010	Checking Customer Control Account	-125,371,902.69
1200	Credit Card Affiliate Vendor Payment	-187,579,360.4
2100	Interbank / External Bank Obligations	374,856,306
2110	Interbank / External Other Withdrawal Obligations	-284,677,638.9
	Sum:	**-222,772,595.99**

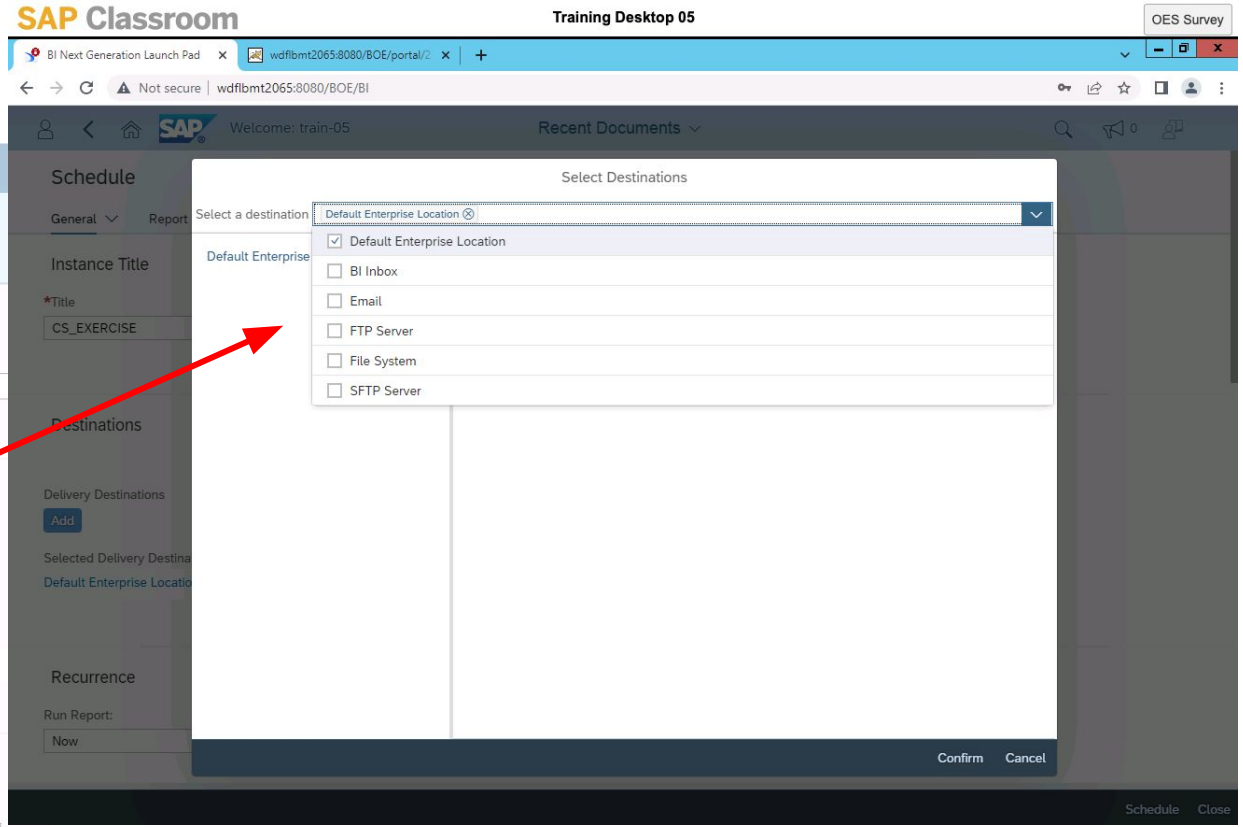


Scheduling

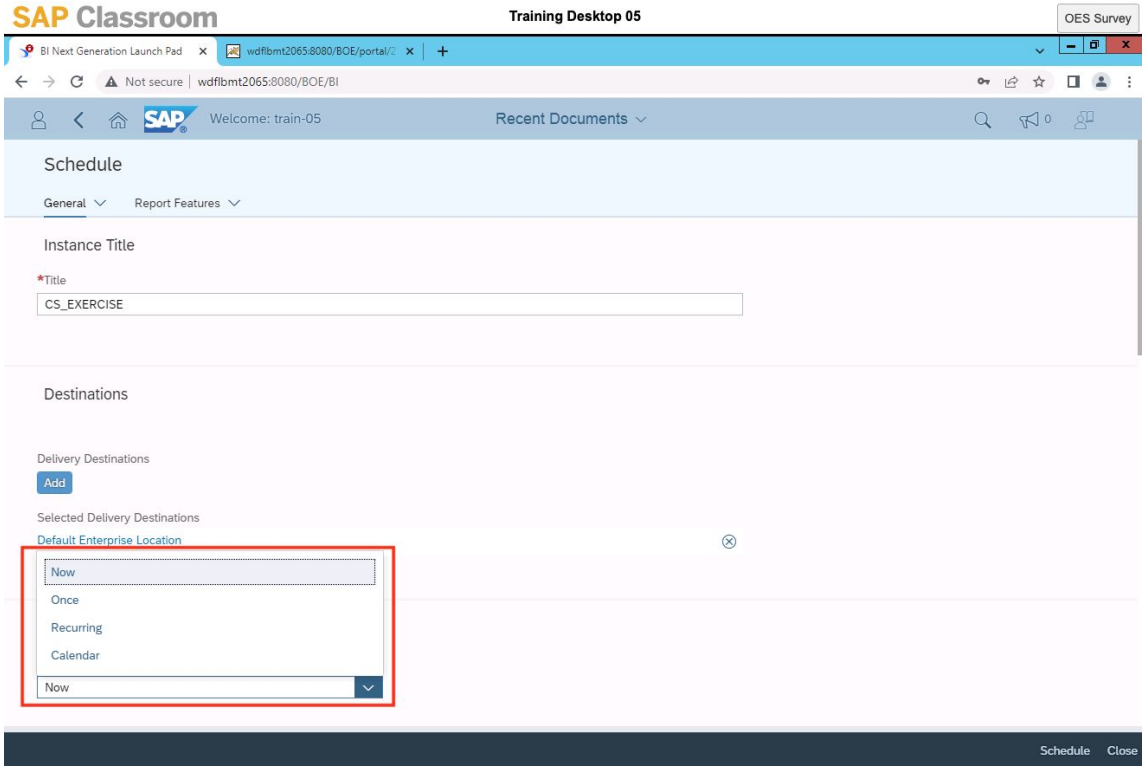

Welcome: train-05
Recent Documents
All Recent Documents (8)
Title
Favorites
Type
Description
Last Updated
CS_EXERCISE
Web Intelligence
Business Analysis on Transactio...
Apr 7, 2022 10:51 AM
CS_EXERCISE_PDF
Web Intelligence
Business Analysis on Transactio...
Mar 17, 2022
CS_EXERCISE_REPORTING
Web Intelligence
Mar 20, 2022
CS_EXERCISE - 2022-03-1...
Microsoft Excel
Business Analysis on Transactio...
Mar 17, 2022
CS_EXERCISE_4
Web Intelligence
Mar 15, 2022
CS_EXERCISE_3
Web Intelligence
Mar 15, 2022
U11 Sol2
Web Intelligence
Mar 12, 2022
Format Breaks and Vertical...
Web Intelligence
Mar 8, 2021
View
View Latest Instance
Properties
Modify
Schedule
Mobile Properties
History
Categories
Mark As Favorite
Details
Send To
Delete

Scheduling

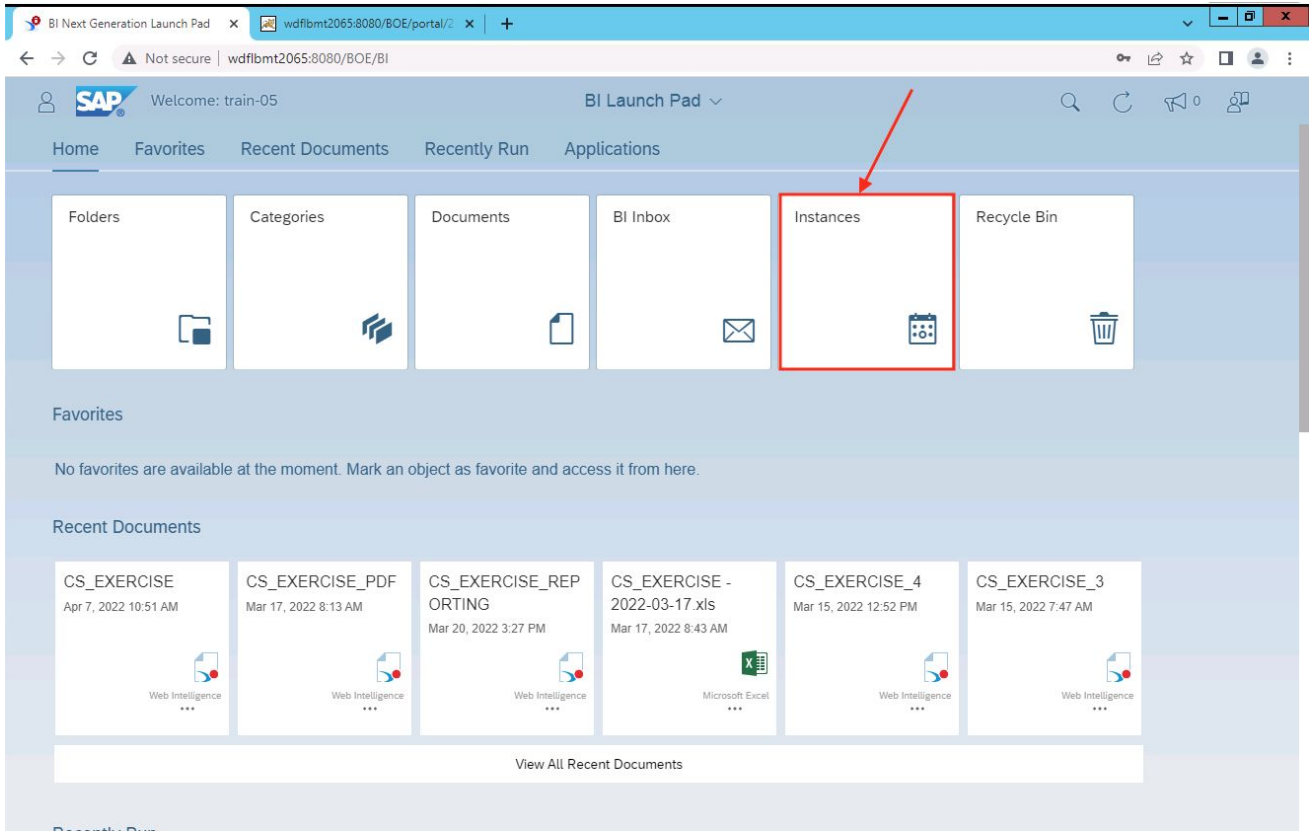

Welcome: train-05
Schedule
General
Report Features
Instance Title
*Title
CS_EXERCISE
Destinations
Delivery Destinations
Add
Selected Delivery Destinations
Default Enterprise Location
Recurrence
Run Report:
Now


SAP Classroom
Training Desktop 05
OES Survey
BI Next Generation Launch Pad
wdflbmt2065:8080/BOE/portal/2
Not secure
wdflbmt2065:8080/BOE/BI
Welcome: train-05
Recent Documents
Schedule
General
Report
Instance Title
*Title
CS_EXERCISE
Destinations
Delivery Destinations
Add
Selected Delivery Destina
Default Enterprise Locatio
Recurrence
Run Report:
Now
Select Destinations
Select a destination
Default Enterprise Location
Default Enterprise Location
BI Inbox
Email
FTP Server
File System
SFTP Server
Confirm
Cancel
Schedule
Close

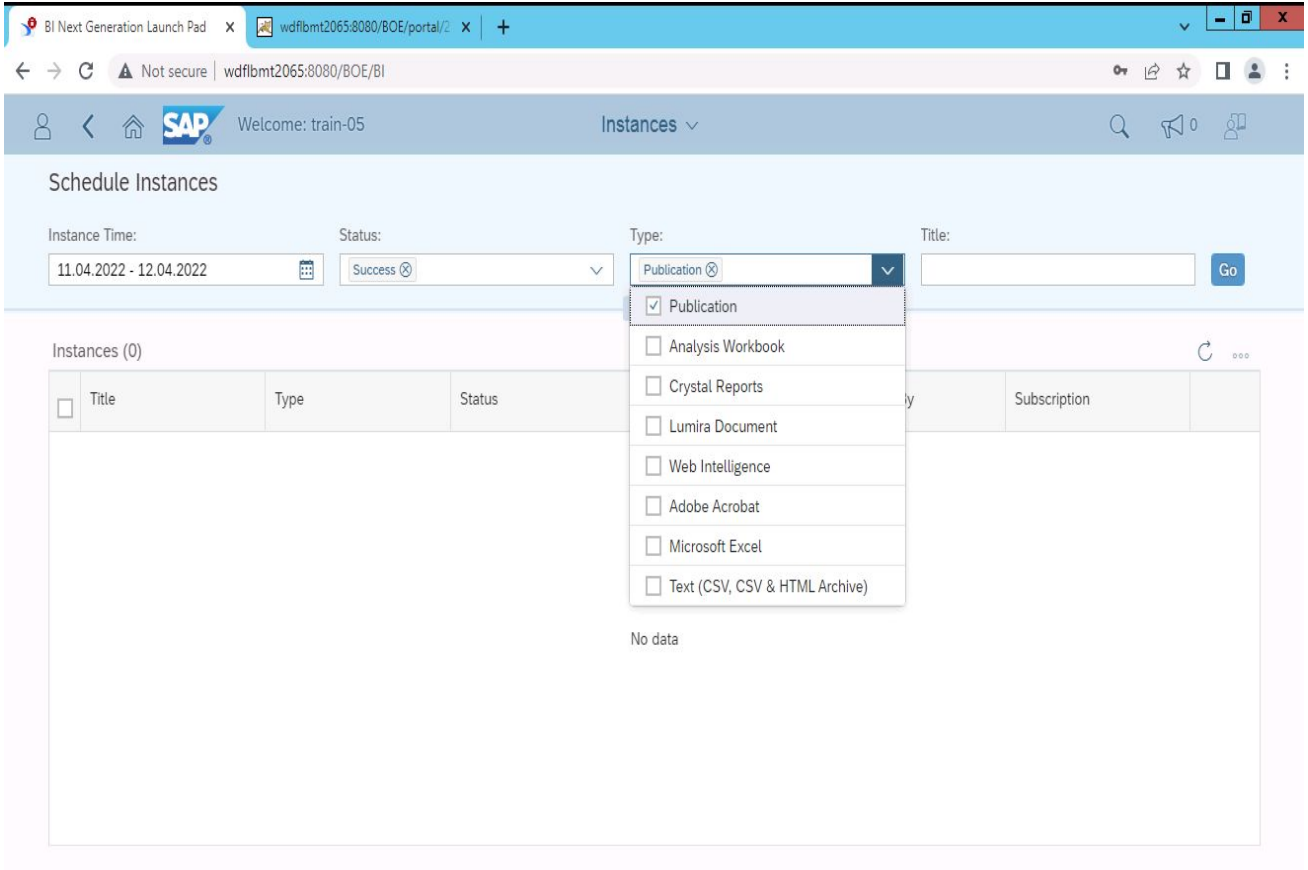
Scheduling

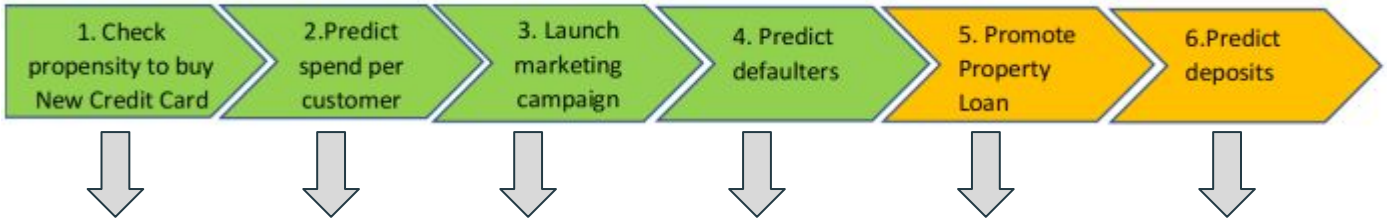

SAP Classroom
Training Desktop 05
OES Survey
BI Next Generation Launch Pad
wdflbmt2065:8080/BOE/portal/2
Not secure
wdflbmt2065:8080/BOE/BI
Welcome: train-05
Recent Documents
Schedule
General
Report Features
Instance Title
*Title
CS_EXERCISE
Destinations
Delivery Destinations
Add
Selected Delivery Destinations
Default Enterprise Location
Now
Once
Recurring
Calendar
Now
Schedule
Close

Publishing


BI Next Generation Launch Pad
wdflbmt2065:8080/BOE/portal/2
Not secure | wdflbmt2065:8080/BOE/BI
Welcome: train-05
BI Launch Pad
Home
Favorites
Recent Documents
Recently Run
Applications
Folders
Categories
Documents
BI Inbox
Instances
Recycle Bin
Favorites
No favorites are available at the moment. Mark an object as favorite and access it from here.
Recent Documents
CS_EXERCISE
Apr 7, 2022 10:51 AM
CS_EXERCISE_PDF
Mar 17, 2022 8:13 AM
CS_EXERCISE_REPORTING
Mar 20, 2022 3:27 PM
CS_EXERCISE - 2022-03-17.xls
Mar 17, 2022 8:43 AM
CS_EXERCISE_4
Mar 15, 2022 12:52 PM
CS_EXERCISE_3
Mar 15, 2022 7:47 AM
Web Intelligence
Microsoft Excel
View All Recent Documents

Publishing


BI Next Generation Launch Pad
wdflbmt2065:8080/BOE/portal/2
Not secure | wdflbmt2065:8080/BOE/BI
Welcome: train-05
Instances
Schedule Instances
Instance Time:
11.04.2022 - 12.04.2022
Status:
Success
Type:
Publication
Title:
Go
Publication
Analysis Workbook
Crystal Reports
Lumira Document
Web Intelligence
Adobe Acrobat
Microsoft Excel
Text (CSV, CSV & HTML Archive)
Instances (0)
Title
Type
Status
Subscription
No data



Classification Model	Regression Model	Clustering Model	Classification Model	Clustering Model	Regression Model

Scenario 1: Propensity To Buy New Credit Card

Business Understanding

What we are doing?

- New Credit card product launch
- Predictive modelling on customers' inclination

End Goals:

- 10% of existing loyal customer to take up the new product.
- Develop personalized products & improve communications

Model Overview

Model: TARGET_CS1_TSP_CC_Propensity	
Data Set:	BANK_AR_2_CS1_TSP_CC_Propensity_1649745796
Analytical Record:	BANK_AR_2
Time-stamped Population:	CS1_TSP_CC_Propensity
Entity:	CARD_ID
Reference Date:	2014-01-01
Initial Number of Variables:	278
Number of Selected Variables:	315
Number of Records:	328,106
Building Date:	2022-04-12 14:55:25
Learning Time:	10mn 54s
Engine Name:	Kxen.RobustRegression
Author:	train-03

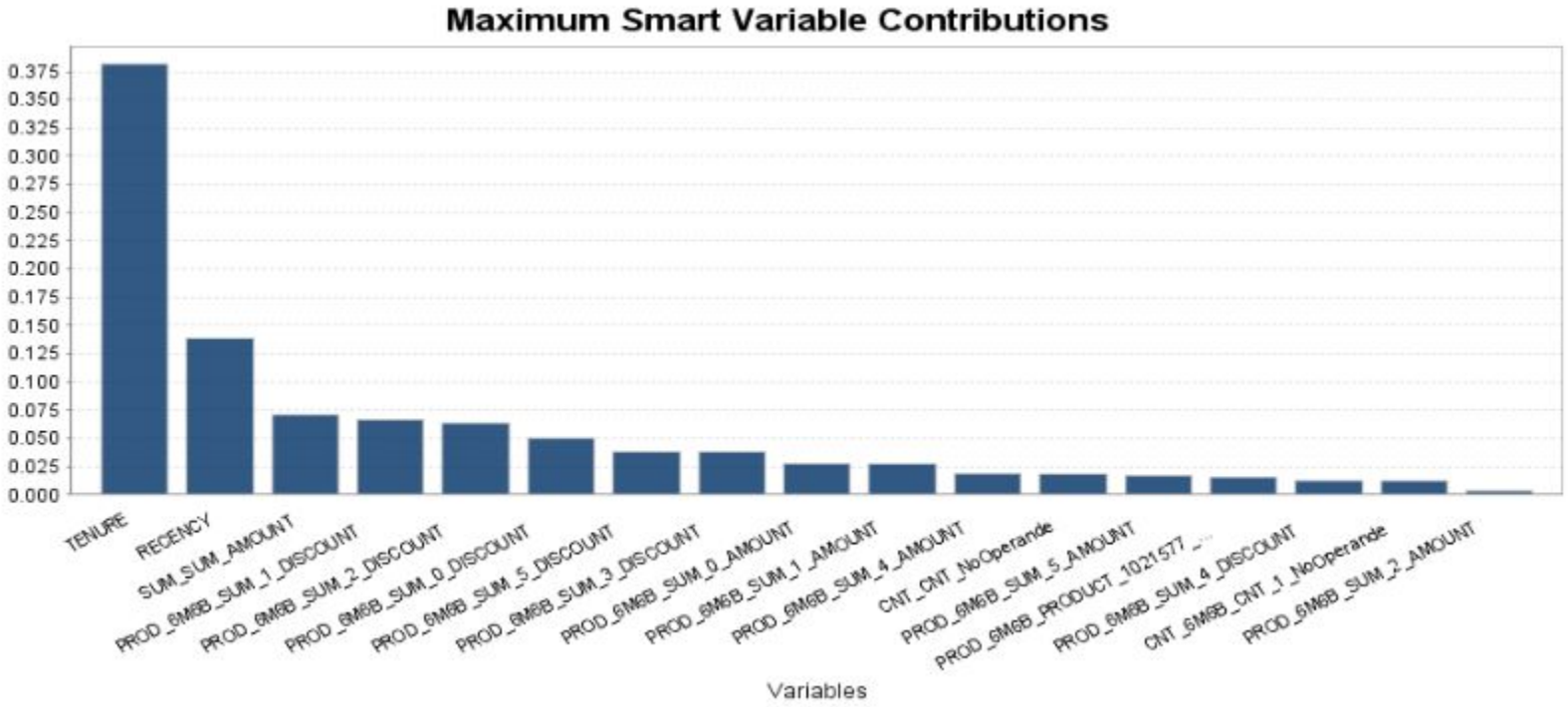
Nominal Targets

TARGET	
Target Key	1
0 - Frequency	97.85%
1 - Frequency	2.15%

Selection Process Selected Iteration

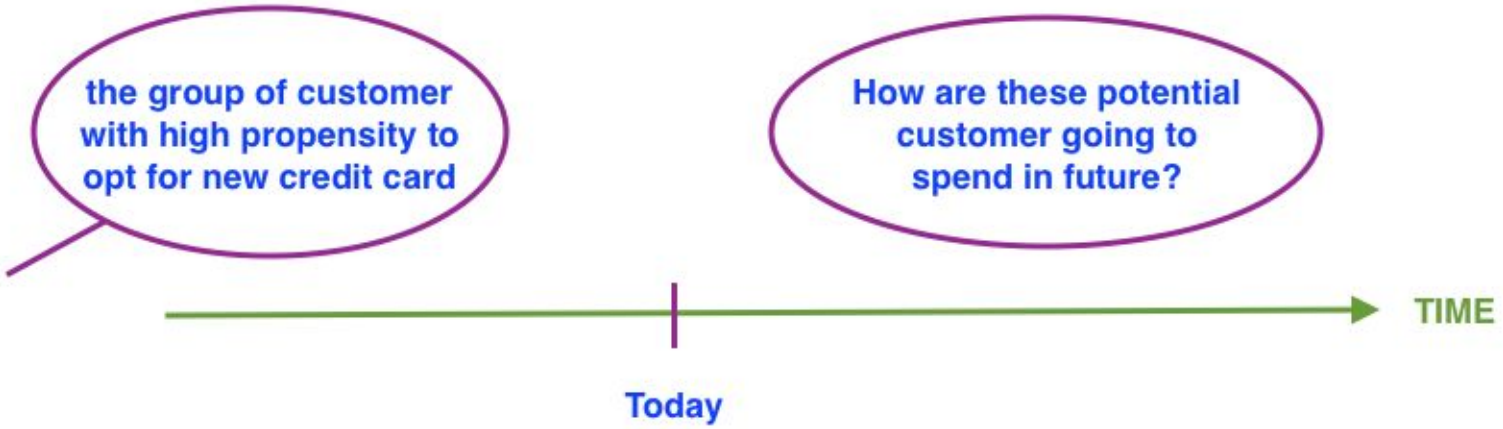
2	
Predictive Power (KI)	0.5076
Prediction Confidence (KR)	0.9814
Nb. Variables Kept	17

Contribution by Variables



Model Graph on Performance

Scenario 2: Predict Customer Spending

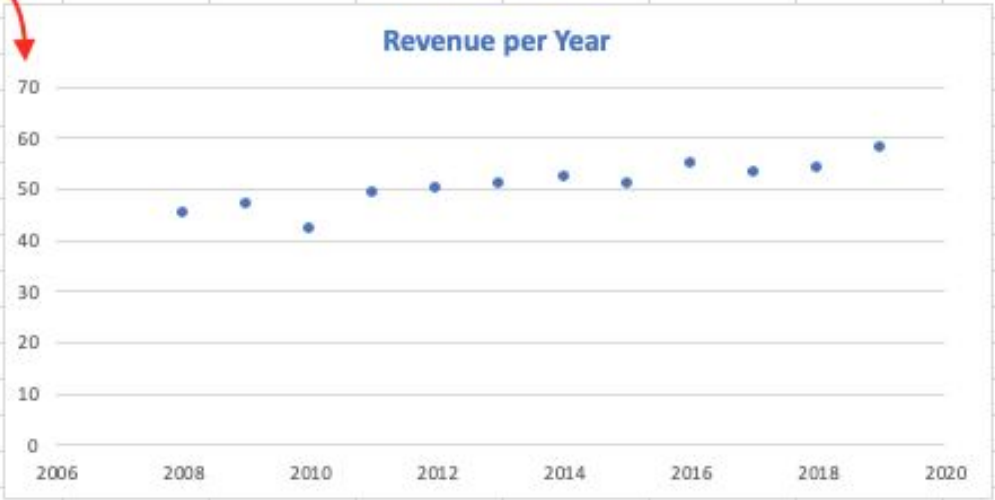

the group of customer with high propensity to opt for new credit card
How are these potential customer going to spend in future?
TIME
Today

Regression Modelling

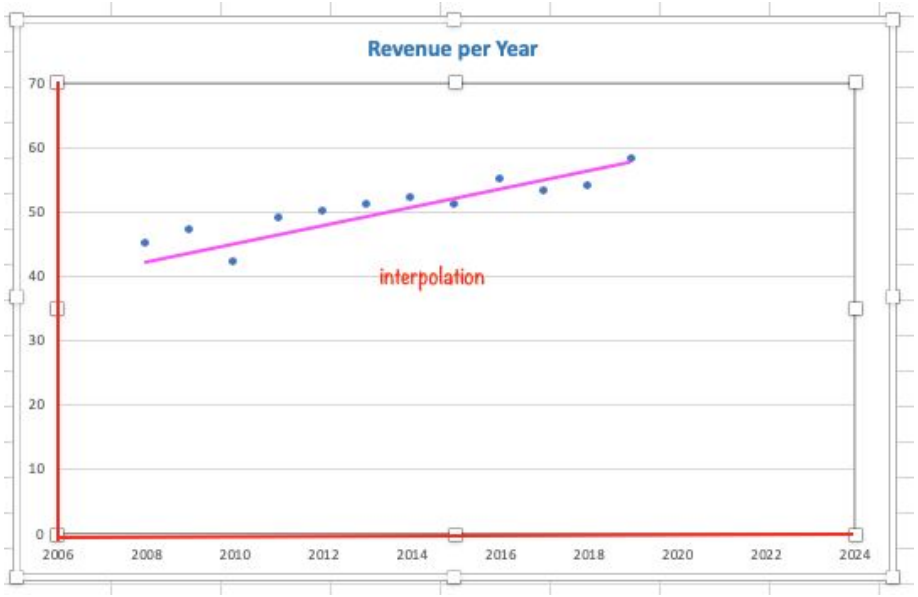
1. Linear
2. Multiple Linear
3. Non-linear

Regression Modelling

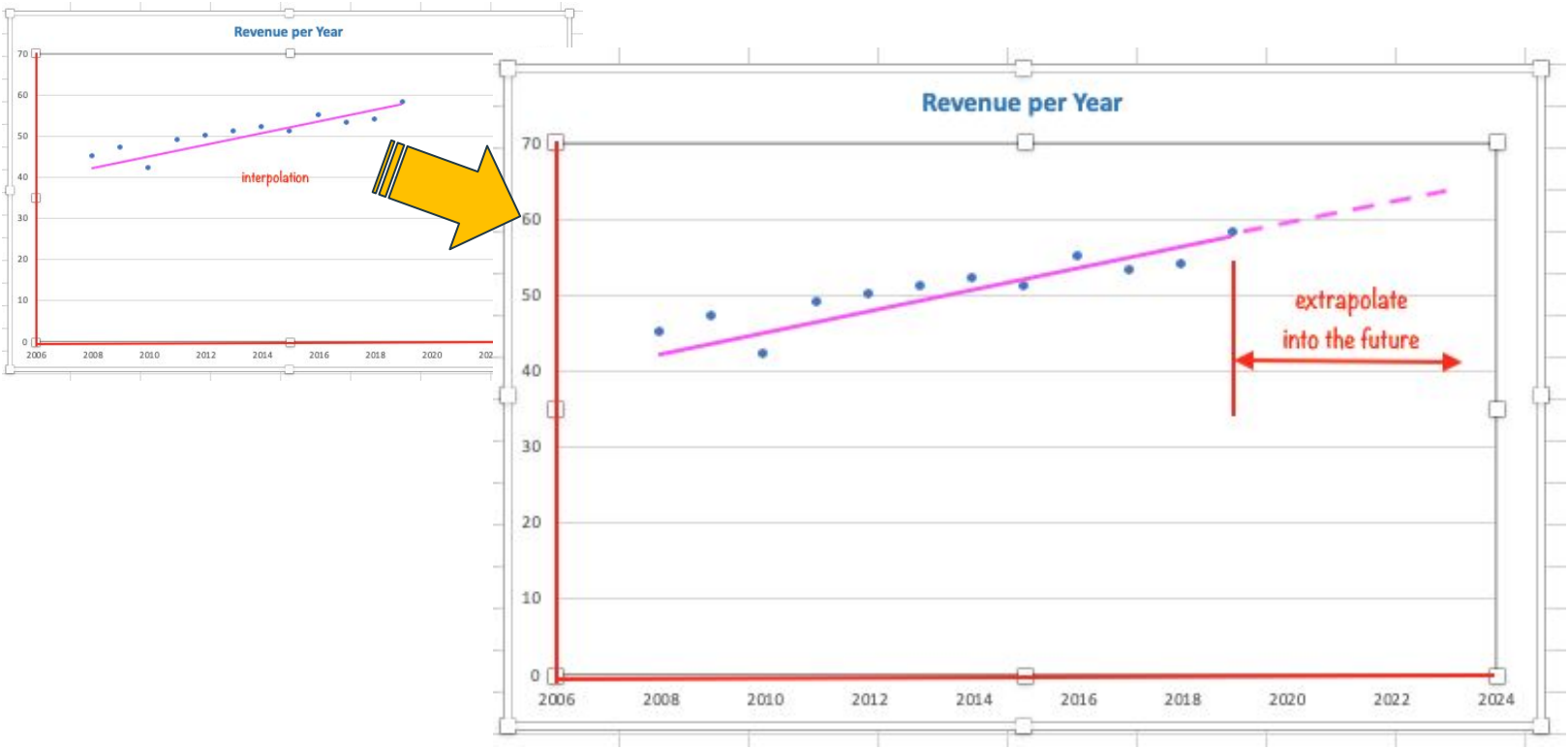
Year	Revenue
2008	45
2009	47
2010	42
2011	49
2012	50
2013	51
2014	52
2015	51
2016	55
2017	53
2018	54
2019	58



Regression Modelling



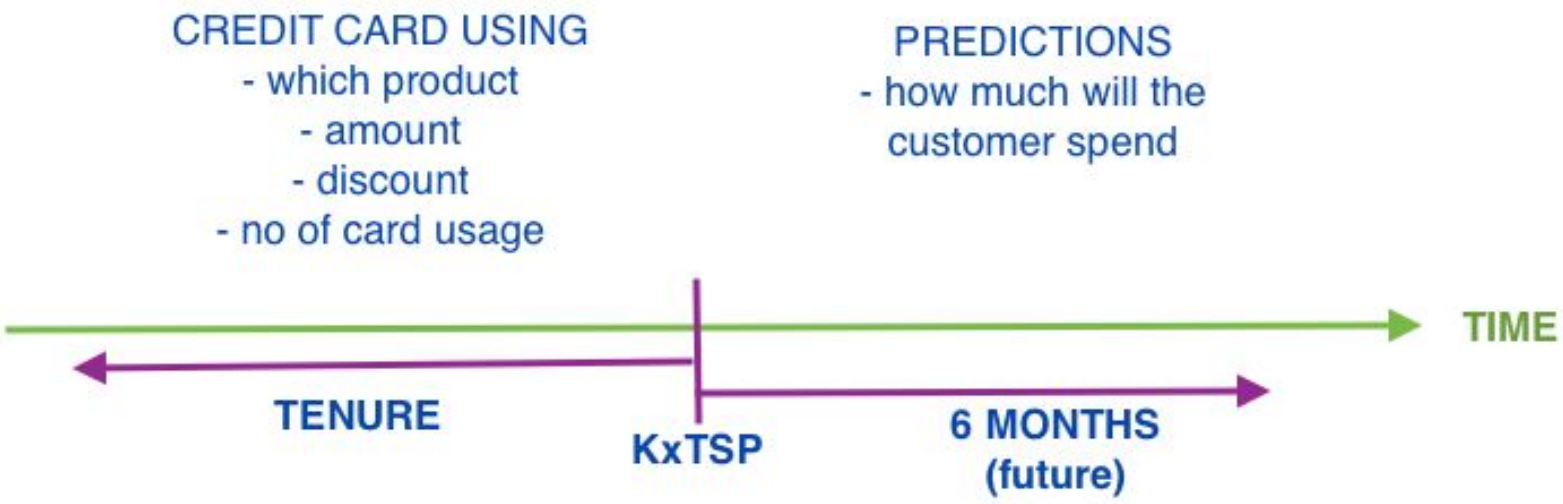
Regression Modelling



Regression Modelling

Business Goal	Data Science Goal
How the prospective customer spend on the credit card in next 6 months once they purchased the new card	KI to strive higher if possible KR above 0.95 to ensure robustness

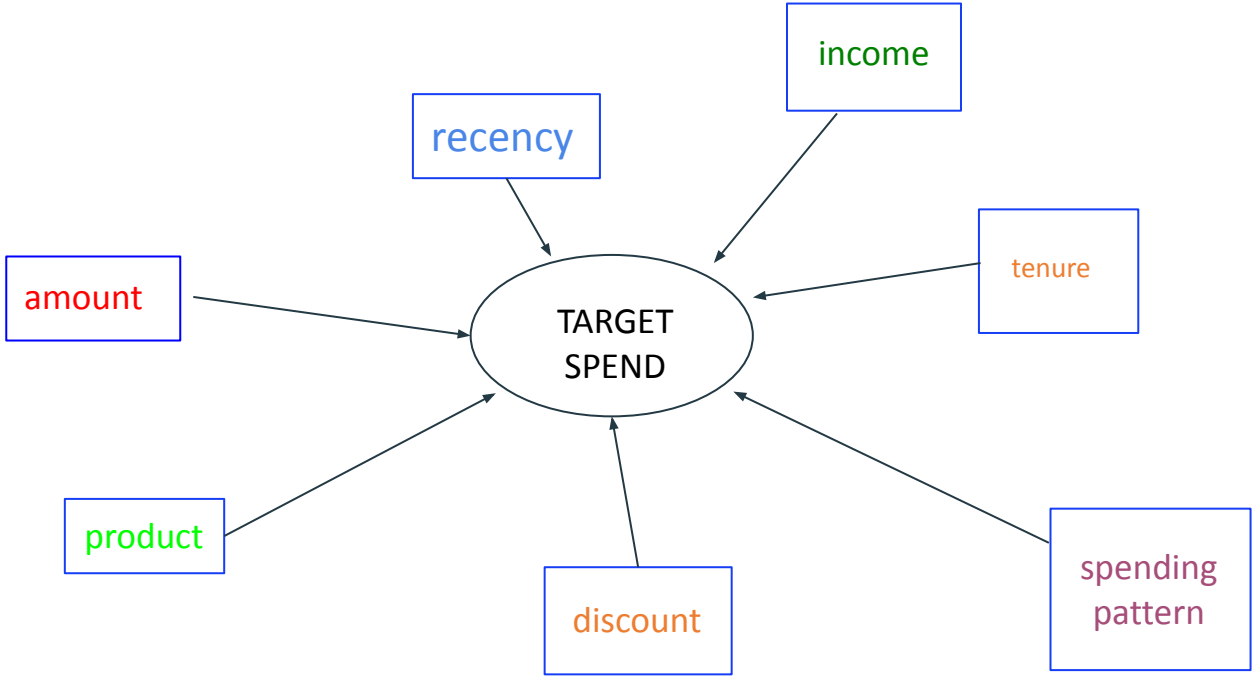
Regression Modelling



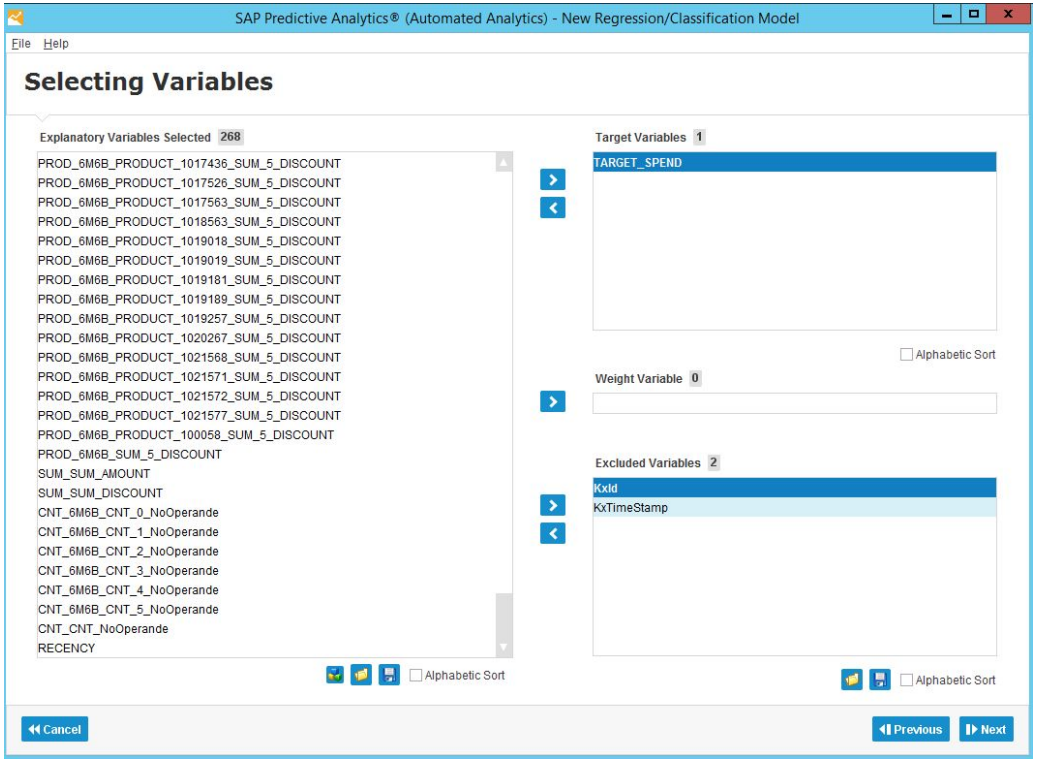
Regression Modelling

TARGET
SPEND

Regression Modelling



Regression Modelling



Regression Modelling

Training the Model

Stop | View Type | Copy | Print | Save | Export to PowerPoint

Report Type: Model Overview

Model Overview

Overview

Model: TARGET_SPEND_TSP Spend in nxt 6 mths

Data Set:	BANK_AR_2_TSP Spend in nxt 6 mths_1649667397
Analytical Record:	BANK_AR_2
Time-stamped Population:	TSP Spend in nxt 6 mths
Entity:	CARD_ID
Reference Date:	2014-01-01
Initial Number of Variables:	259
Number of Selected Variables:	256
Number of Records:	248,078
Building Date:	2022-04-11 17:09:57
Learning Time:	11mn 57s
Engine Name:	Kxen.RobustRegression
Author:	train-05

Continuous Targets (Number)

TARGET_SPEND

Min	0
Max	262,019
Mean	912.984
Standard Deviation	2,248.66

Selection Process Selected Iteration

2

Predictive Power (KI)	0.4793
Prediction Confidence (KR)	0.9958
Nb. Variables Kept	18

Cancel | Previous | Next

KI = 0.4793

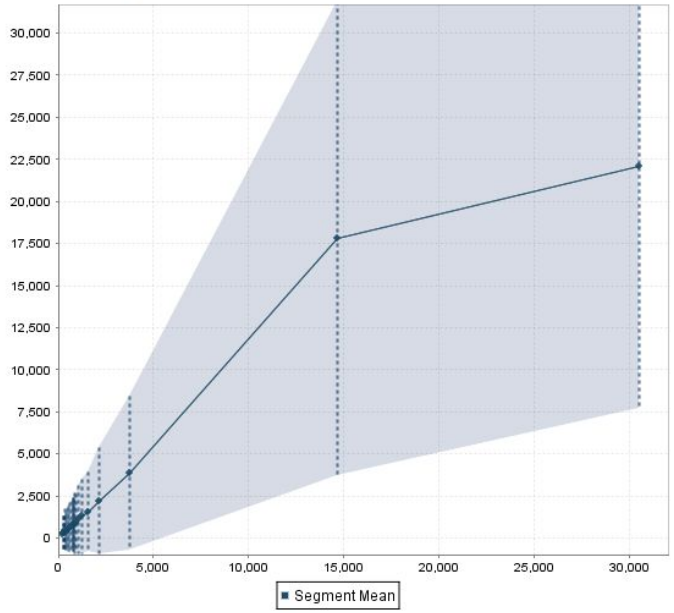
KR = 0.9958

No of Variable kept = 18

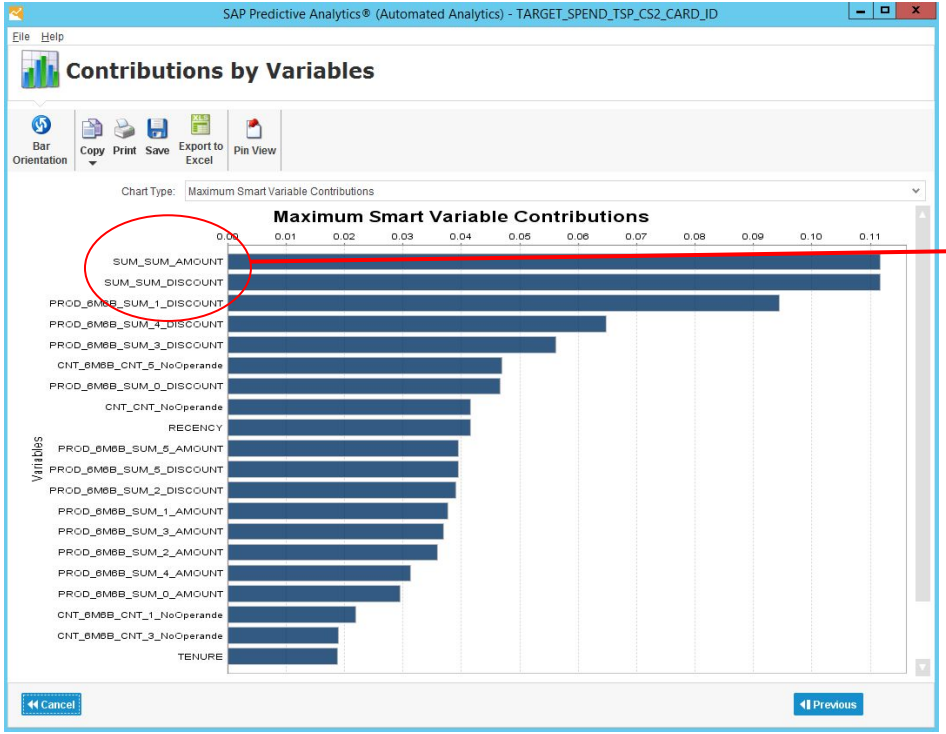
Regression Modelling

Report Type: Executive Report

Estimator: rr_TARGET_SPEND, Target: TARGET_SPEND



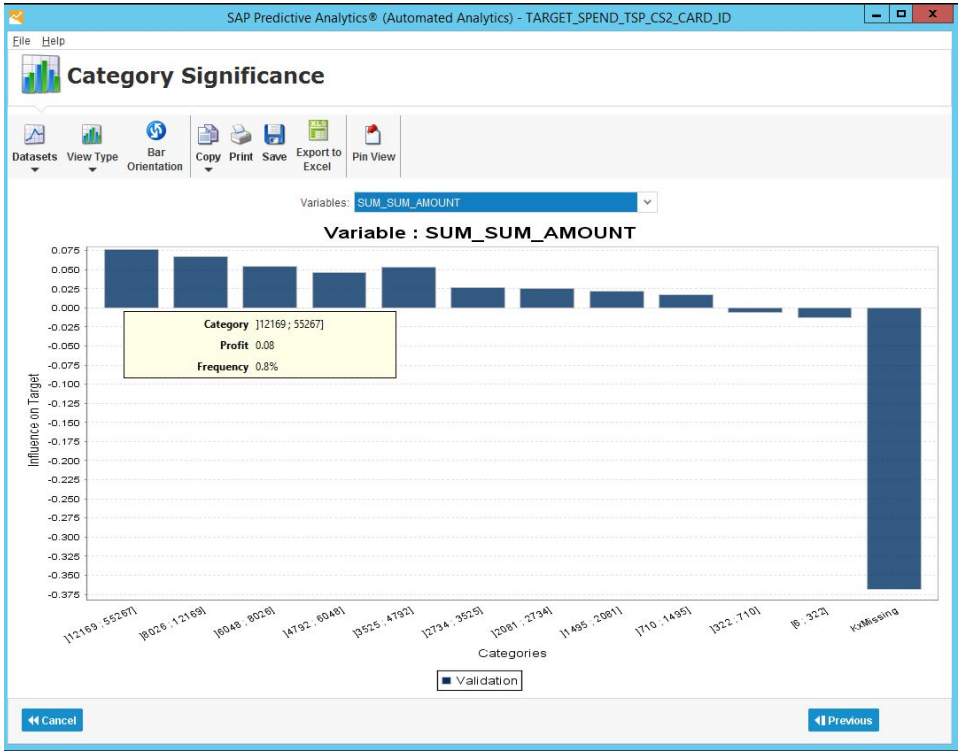
Regression Modelling



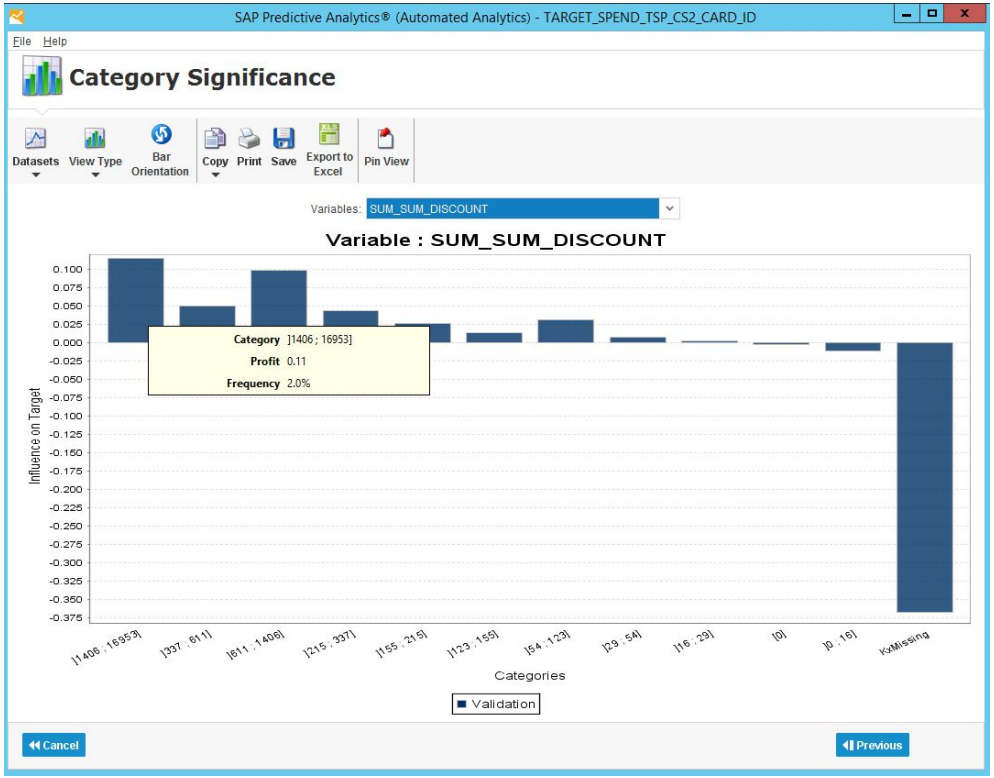
Amount

Discount

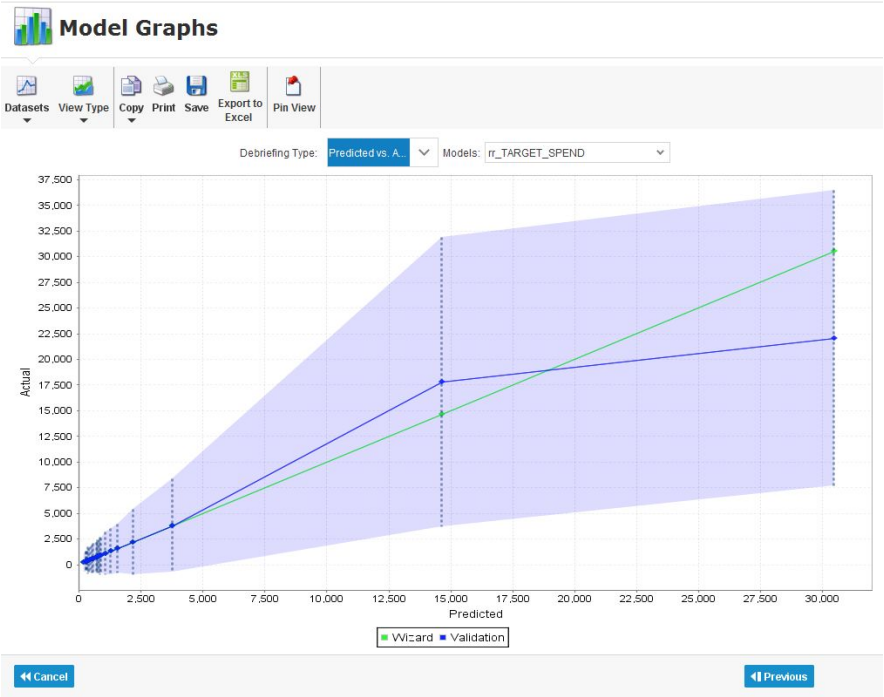
Regression Modelling



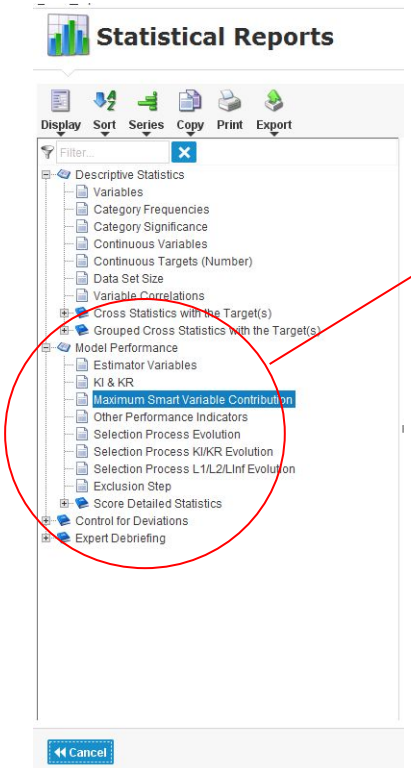
Regression Modelling

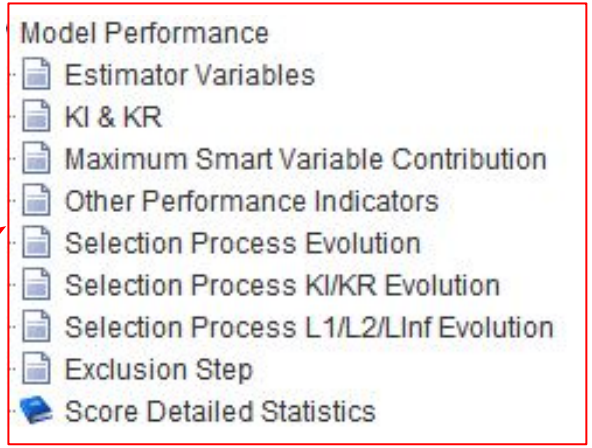


Regression Modelling



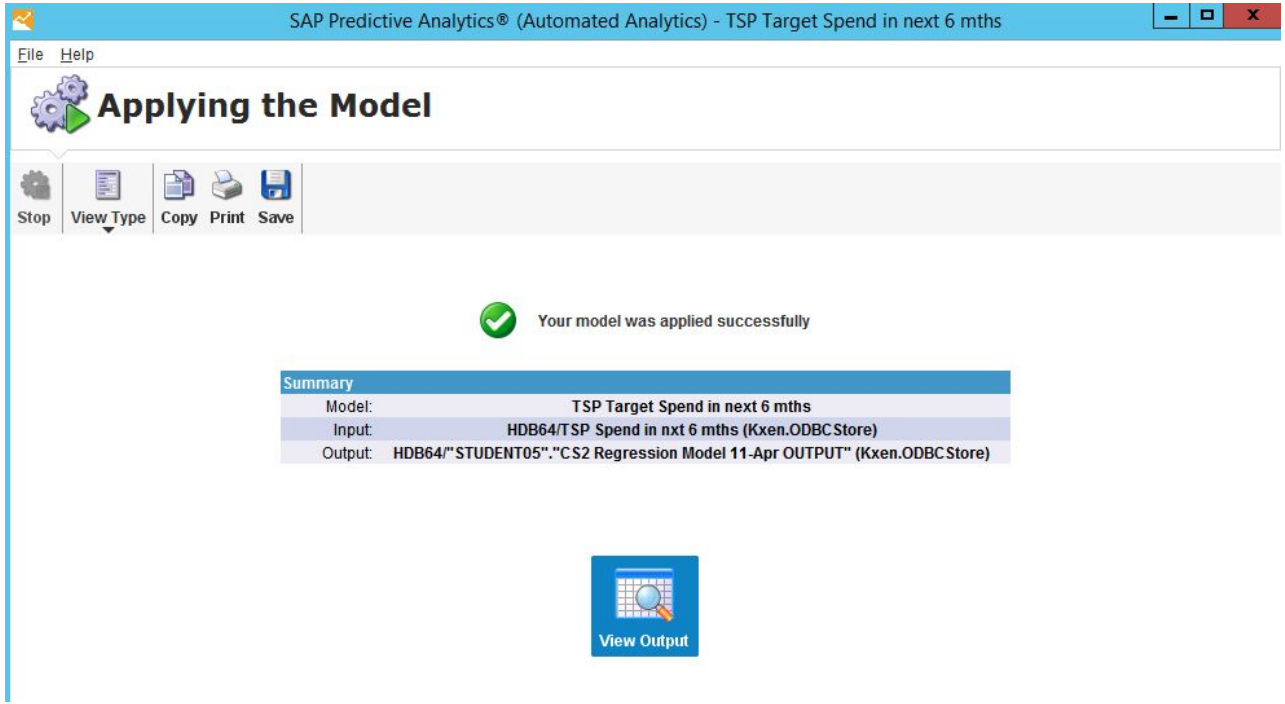
Regression Modelling





Indicator	Estimation	Test	Validation
L1	1,104.19	1,089.64	1,101.82
L2	1,938.7	1,823.17	1,911.36
LInf	157,099	35,250	50,896
ErrorMean	6.286	-16.426	1.495
ErrorStdDev	1,938.69	1,823.1	1,911.36
R2	0.261	0.241	0.213

Regression Modelling - Application



Regression Modelling - Application

File Help

Applying the Model

Stop | View Type | Copy | Print | Save

Output Name: "STUDENT05".CS2 Regression Model 11-Apr OUTPUT"

First Row Index: 1 Last Row Index: 2,000

Data | Statistics | Graph

	KxId	KxTimeStamp	TARGET_SPEND	rr_TARGET_SPEND
8	69741104	2014-07-01	18537	1524
9	71167732	2014-07-01	2867	1233
10	71313436	2014-07-01	3031	399
11	71907284	2014-07-01	0	384
12	71907536	2014-07-01	0	383
13	72204160	2014-07-01	0	1006
14	73064084	2014-07-01	0	866
15	73081920	2014-07-01	0	465
16	73129172	2014-07-01	0	394
17	73175032	2014-07-01	515	383
18	73183136	2014-07-01	0	408
19	73547976	2014-07-01	1940	557
20	73554516	2014-07-01	0	383
21	73557784	2014-07-01	0	399
22	73698524	2014-07-01	0	399
23	73700500	2014-07-01	2491	556
24	73839996	2014-07-01	1186	1491
25	74069200	2014-07-01	0	701
26	74079452	2014-07-01	0	895
27	74093304	2014-07-01	0	470
28	74206636	2014-07-01	3051	1812
29	74270088	2014-07-01	0	383

Cancel | Previous | Next

Scenario 3: Launch Marketing Campaign

Scenario 3: Launch Marketing Campaign

Objectives

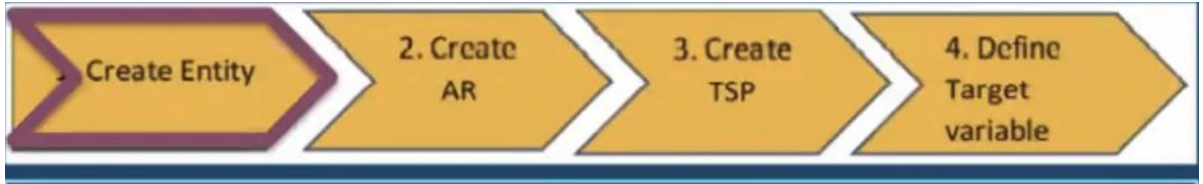
1. Given the large customer base, the Bank would like to develop group specific advertising strategy so that each advertisement will be relevant to the targeted group.
2. The bank would like to develop 3 to 5 such advertisements.
3. The Bank has transaction data for a twelve-month period from January to December 2014.

Discussion Points

1. What model will you use for above requirement?
   - **Cluster model** will be used since we would like to create marketing campaign for a specific group of customers.
2. Will you choose supervised or unsupervised model?
   - Since the Bank has informed that they would like to create 3 to 5 marketing campaigns, we already know the number of clusters. Hence this is a **supervised model**.
3. Which other tool can be used to develop clustering?
   - **Only Predictive Analytics offer clustering**. SAP Predictive Factory does not offer cluster modelling. However, the model built in Predictive Analytics can be imported into Predictive Factory.

1) Prepare Time Stamped Population

Analytical Dataset [ADS] Preparation



Entity (Same as scenario 1 & 2):

Unique identifier CARD_ID from CS_CARD_TRANSACTIONS

- Name = CARD_ID
- Description = Card ID
- Table = /Catalog/TRAINING/CS_LOYALTY_CARD
- Id Field = CARD (Card ID number)

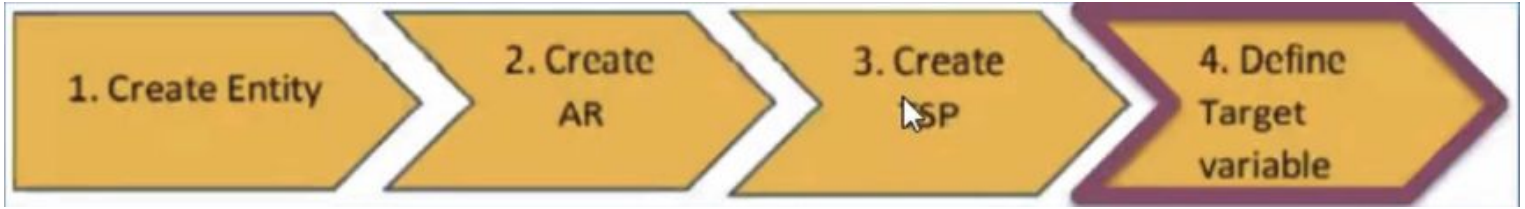
1) Prepare Time Stamped Population

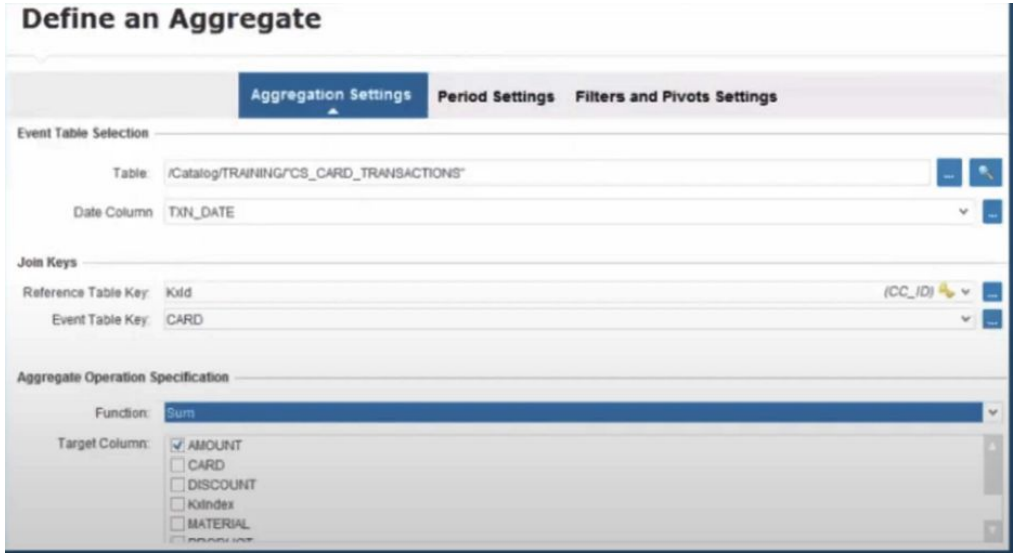
Analytical Dataset [ADS] Preparation

	Alias	Name	Type	Storage	Source	Visibility	Order	Key	Description
254	PROD_6M6B_PR...	PROD_6M6B_PR...	continuous	Integer	Aggregate	✓	0	0	Sum(DISCOUNT)...
255	PROD_6M6B_PR...	PROD_6M6B_PR...	continuous	Integer	Aggregate	✓	0	0	Sum(DISCOUNT)...
256	PROD_6M6B_PR...	PROD_6M6B_PR...	continuous	Integer	Aggregate	✓	0	0	Sum(DISCOUNT)...
257	PROD_6M6B_PR...	PROD_6M6B_PR...	continuous	Integer	Aggregate	✓	0	0	Sum(DISCOUNT)...
258	PROD_6M6B_PR...	PROD_6M6B_PR...	continuous	Integer	Aggregate	✓	0	0	Sum(DISCOUNT)...
259	PROD_6M6B_SU...	PROD_6M6B_SU...	continuous	Integer	Aggregate	✓	0	0	Sum(DISCOUNT)...
260	SUM_AVG_AMOU...	SUM_AVG_AMOU...	continuous	Number	Aggregate	✓	0	0	Average(AMOUNT)
261	SUM_AVG_DISC...	SUM_AVG_DISC...	continuous	Number	Aggregate	✓	0	0	Average(DISCOU...
262	CNT_6M6B_CNT...	CNT_6M6B_CNT...	continuous	Integer	Aggregate	✓	0	0	Count(AMOUNT) | ...
263	CNT_6M6B_CNT...	CNT_6M6B_CNT...	continuous	Integer	Aggregate	✓	0	0	Count(AMOUNT) | ...
264	CNT_6M6B_CNT...	CNT_6M6B_CNT...	continuous	Integer	Aggregate	✓	0	0	Count(AMOUNT) | ...
265	CNT_6M6B_CNT...	CNT_6M6B_CNT...	continuous	Integer	Aggregate	✓	0	0	Count(AMOUNT) | ...
266	CNT_6M6B_CNT...	CNT_6M6B_CNT...	continuous	Integer	Aggregate	✓	0	0	Count(AMOUNT) | ...
267	CNT_6M6B_CNT...	CNT_6M6B_CNT...	continuous	Integer	Aggregate	✓	0	0	Count(AMOUNT) | ...
268	CNT_CNT_NoOp...	CNT_CNT_NoOp...	continuous	Integer	Aggregate	✓	0	0	Count(Kxen.Aggr...
270	RECENCY	RECENCY	continuous	Integer	Function	✓	0	0	
	Kxid	Kxid	nominal	Integer	CC_IDPopulation	✓	0	1	
272	KxTimeStamp	KxTimeStamp	nominal	Date	CC_IDPopulation	✓	0	1	
274	GENDER_TYPE	GENDER_TYPE	nominal	String	CatalogTRAININ...	✓	0	0	
275	AGE_GROUP	AGE_GROUP	nominal	String	CatalogTRAININ...	✓	0	0	
276	EMAIL_FLG	EMAIL_FLG	nominal	String	CatalogTRAININ...	✓	0	0	
277	ADDRESS_FLG	ADDRESS_FLG	nominal	String	CatalogTRAININ...	✓	0	0	

1) Prepare Time Stamped Population

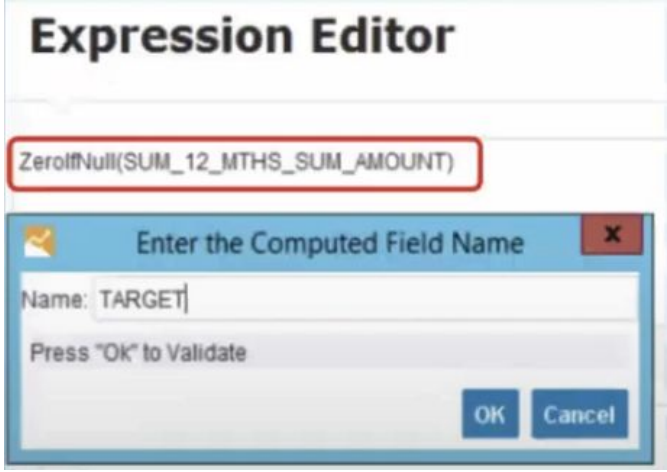
Define and Aggregate



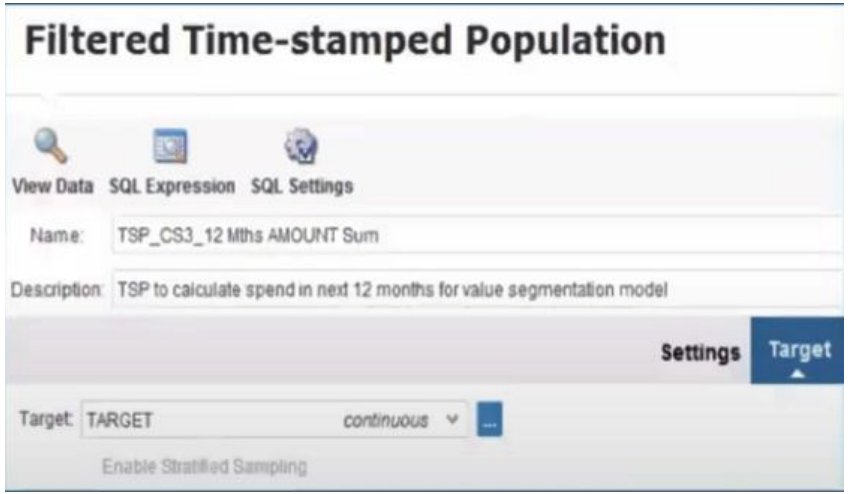


1) Prepare Time Stamped Population

Target Variable: SUM_12_MTHS



Time Stamp Population:



2) Build Cluster Model

Overview

Model Overview

Overview

Model: TARGET_TSP_CS3_12 Mths AMOUNT Sum

Data Set:	BANK_AR_2_TSP_CS3_12 Mths AMOUNT Sum_1634809773
Analytical Record:	BANK_AR_2
Time-stamped Population:	TSP_CS3_12 Mths AMOUNT Sum
Entity:	CC_ID
Reference Date:	2014-01-01
Initial Number of Variables:	277
Number of Selected Variables:	326
Number of Records:	19,401
Building Date:	2021-10-21 17:57:09
Learning Time:	6mn 53s
Engine Name:	Kxen.SmartSegmenter
Author:	train-05
Minimum Requested Number of Clusters:	3
Maximum Requested Number of Clusters:	5
SQL Expressions:	enabled

Continuous Targets (Number)

TARGET

Min	0
Max	73,669
Mean	1,453.08
Standard Deviation	3,524.53

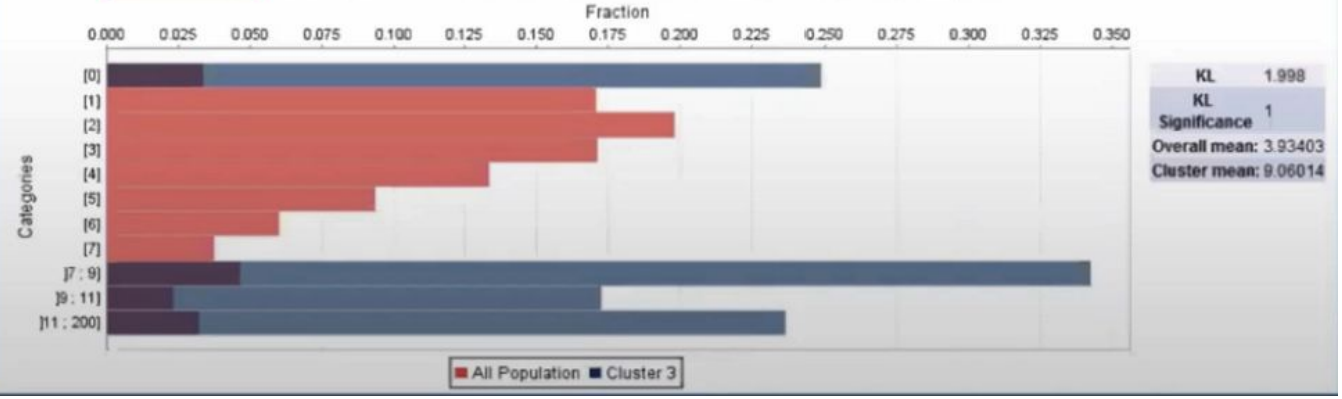
2) Build Cluster Model

"**Cluster 3**" Profile

Cluster Name	Frequencies	Target Means
1	14.49%	1559.8
2	36.54%	858.881
3	13.57%	3004.16
4	14.22%	1021.6
5	15.99%	1764.77

Variables: CNT_CNT_NoOperande ... Freeze Variable

"Cluster 3" vs. Whole Population for Variable "CNT_CNT_NoOperande"

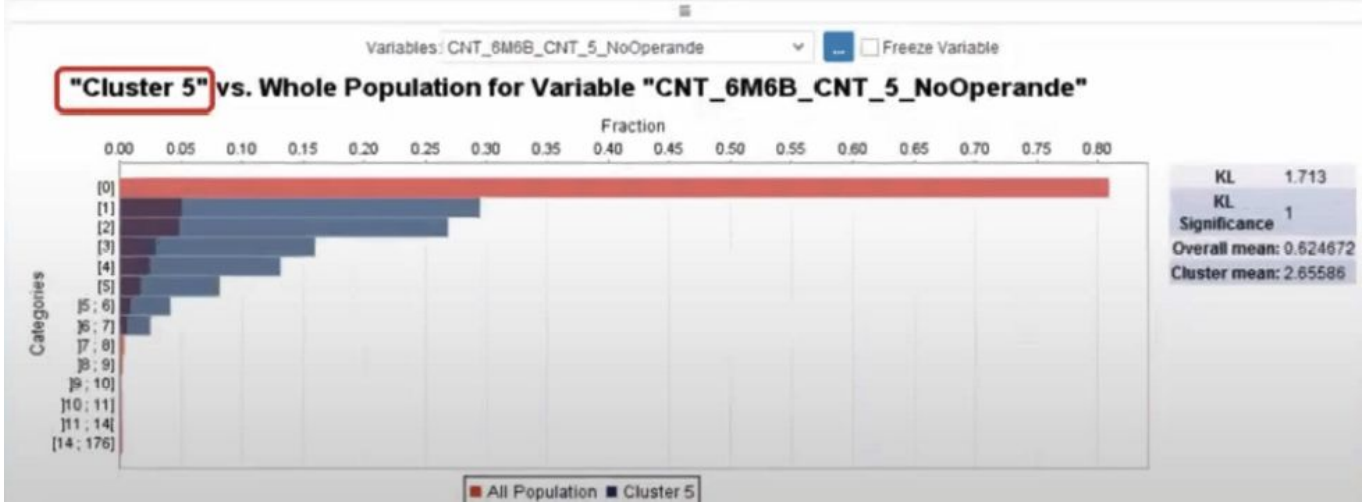


KL	1.998
KL Significance	1
Overall mean:	3.93403
Cluster mean:	9.06014

2) Build Cluster Model

"**Cluster 5**" Profile

Cluster Name	Frequencies	Target Means
1	14.49%	1559.8
2	36.54%	858.881
3	13.57%	3004.16
4	14.22%	1021.6
5	15.99%	1764.77



3) Analyze and Save the Model

Statistical Reports

Cluster Statistics for Continuous Variable(s)

Cluster Variable: kc_TARGET

Data Set: Validation

Variable: TARGET

Cluster	Minimum	Maximum	Mean	Standard Deviation
Non spender Oct 13	0	19724	1,269.62	2,479.39
Last use 95 days ago	0	25221	804.759	2,090.3
High usage count	0	158663	3,356.6	8,382.3
Last use 60 to 94 days ago	0	28318	1,286.64	2,891.25
Non spender Nov 13	0	24113	1,794.02	3,271.11
6	0	15462	1,284.75	2,478.39

Cluster Frequency, Target Mean and Standard Deviation

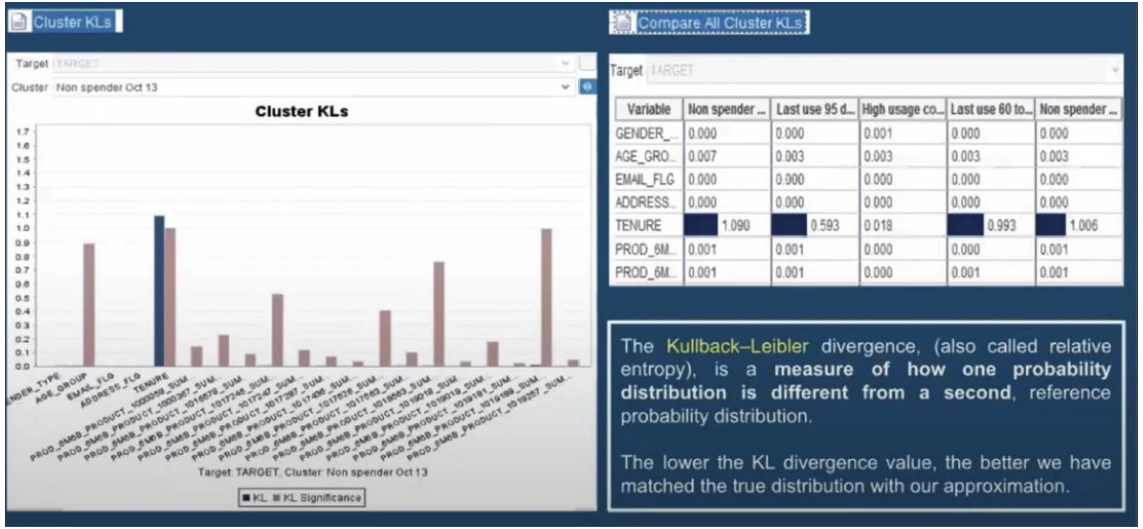
Cluster	Frequency	Target Mean	Target Standard Deviation
Non spender Oct 13	14.49%	1,559.8	3,196.86
Last use 95 days ago	36.54%	858.881	2,126.68
High usage count	13.57%	3,004.16	6,098.13
Last use 60 to 94 days ago	14.22%	1,021.6	2,324.28
Non spender Nov 13	15.99%	1,764.77	3,908.41

Cluster Counts

Target	Initial Number of Clusters	Final Number of Clusters	Overlap	Percentage of Unassigned Records
TARGET	5	5	16.19%	5.2%

3) Analyze and Save the Model

Statistical Reports

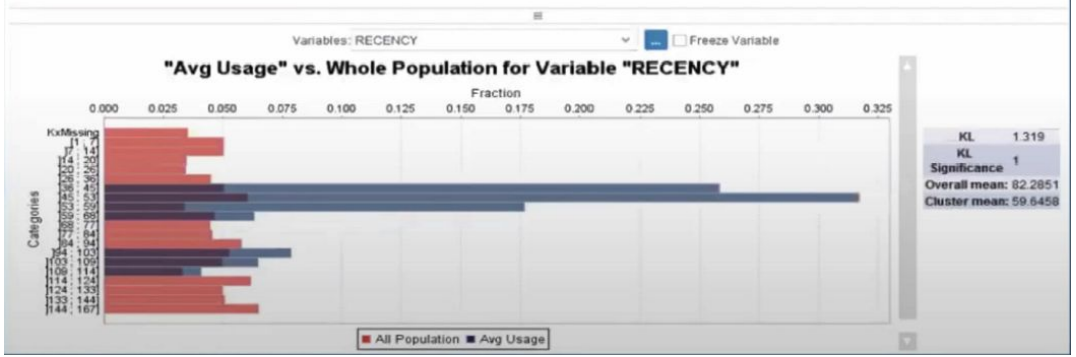


Variable	Non spender ...	Last use 95 d...	High usage co...	Last use 60 to...	Non spender ...
GENDER_...	0.000	0.000	0.001	0.000	0.000
AGE_GRO...	0.007	0.003	0.003	0.003	0.003
EMAIL_FLG	0.000	0.000	0.000	0.000	0.000
ADDRESS...	0.000	0.000	0.000	0.000	0.000
TENURE	1.090	0.593	0.018	0.993	1.006
PROD_6M...	0.001	0.001	0.000	0.000	0.001
PROD_6M...	0.001	0.001	0.000	0.001	0.001

3) Analyze and Save the Model

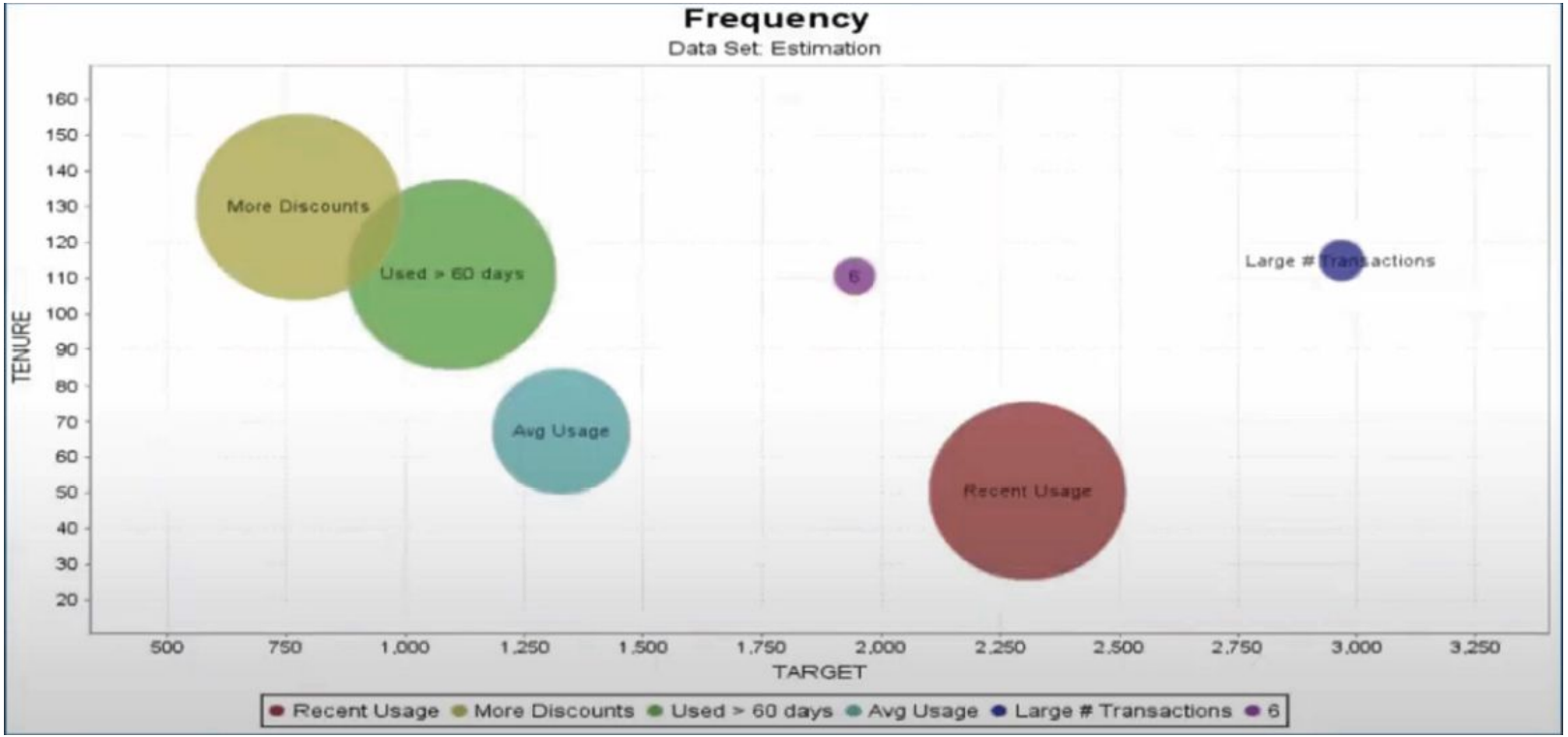
Choice of Model: **Run 1** even though lower KI, but Clusters are more meaningful.

Cluster Name	Frequencies	Target Means
Recent Usage	24.98%	2305.92
More Discounts	26.18%	775.236
Used > 60 days	26.62%	1098.45
Avg Usage	16.5%	1327.9
Large # Transactions	3.18%	2967.6



3) Analyze and Save the Model

Cluster Summary



4) Load the model into SAP Predictive Factory

Data Science Success Criteria

Data Science Success Criteria

KI: as high as possible (not if not meaningful)

KR: > 0.95 (0.9836)

Target	Within Cluster Variance	KI	KR
TARGET	31,267,100	0.2863	0.9836

Low cluster overlap percentage(13.47%).

Run 2: 16.19%

Target	Initial Number of Clusters	Final Number of Clusters	Overlap	Percentage of Unassigned Records
TARGET	5	5	13.47%	2.53%

Low percentage of unassigned records (2.53%).

Run 2: 5.2%

Target	Percentage of Unassigned Records
TARGET	2.53%

Scenario 4:Predict Loan Defaulters

Goals and Success Criteria (POV)

Business Goals	**Business Success Criteria**	**Data Science Goals**	**Data Science Success Criteria**
This project has following objectives: ● Increase profits by identifying performing payments ● Reduce risk by avoiding making card offers to potential defaulters	● The Bank reduces the percentage of Bad Loans from 11 % to 7%	● Build a classification model that predicts Loan Defaulters ● Latency period:nil ● History period:6 mths ● Target period:6 mths ● Population Filters: filter out only those customers who have an outstanding loan ● Target variable:find whether a customer will default on payment as Yes/No	This project will be judged a success if: ● KI>0.45 ● KR>0.95 ● Variable Contributions make business sense ● Model performance in evaluation period corresponds to performance in training period

Analytical Data Set Preparation

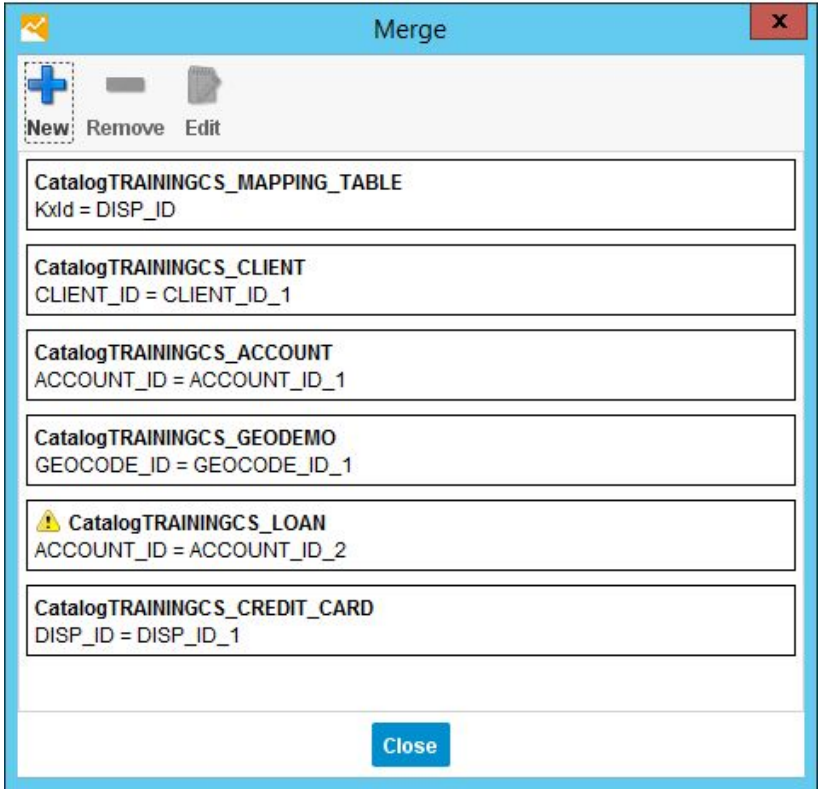
Entity:
Unique Identifier DISP_ID from CS_MAPPING_TABLE

AR:
1.CS_MAPPING_TABLE
2.CS_CLIENT
3.CS_ACCOUNT
4.CS_GEODEMO
5.CS_LOAN
6.CS_CREDIT_CARD

TSP:
Filter out IDs which have outstanding. Outstanding amount can be found in CS_LOAN. Hence merge table CS_LOAN

Target Variable:
Find whether a customer will default on payment as Yes or No

Analytical Record:Merge Tables

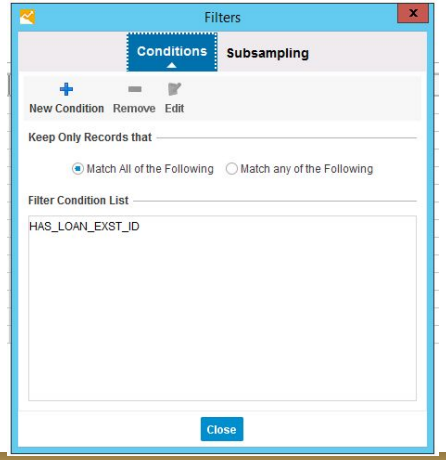


Analytical Record

	Alias	Name	Type	Storage	Source	Visibility	Order	Key	Description
2	AVERAGE_BALANCE_1M6...	AVERAGE_BALANCE_1M6...	continuous	Number	Aggregate	☑	0	0	Average(BALANCE) | 1M6B...
3	AVERAGE_PAYMENT_1M6...	AVERAGE_PAYMENT_1M6...	continuous	Number	Aggregate	☑	0	0	Average(BALANCE) | 1M6B...
4	MIN_PAYMENT_1M6B_MIN...	MIN_PAYMENT_1M6B_MIN...	continuous	Number	Aggregate	☑	0	0	Min(BALANCE) | 1M6B(0/1)
5	MAX_PAYMENT_1M6B_MA...	MAX_PAYMENT_1M6B_MA...	continuous	Number	Aggregate	☑	0	0	Max(BALANCE) | 1M6B(0/1)
7	EST_ANNUAL_INCOME_1...	EST_ANNUAL_INCOME_1...	continuous	Number	Aggregate	☑	0	0	Sum(AMOUNT) | 1Y1B(0/1)
8	CLIENT_AGE	CLIENT_AGE	continuous	Number	Function	☑	0	0	
9	ACCOUNT_AGE	ACCOUNT_AGE	continuous	Integer	Function	☑	0	0	
10	CARD_AGE	CARD_AGE	continuous	Integer	Function	☑	0	0	
... 🔒	KxId	KxId	nominal	Integer	DISP_IDPopulation	☑	0	1	
12	KxTimeStamp	KxTimeStamp	nominal	Date	DISP_IDPopulation	☑	0	1	
14	CLIENT_ID	CLIENT_ID	continuous	Integer	CatalogTRAININGCS_MAP...	☑	0	0	
15	ACCOUNT_ID	ACCOUNT_ID	continuous	Integer	CatalogTRAININGCS_MAP...	☑	0	0	
16	CLIENT_TYPE	CLIENT_TYPE	nominal	String	CatalogTRAININGCS_MAP...	☑	0	0	
18	HOME_GEOCODE_ID	HOME_GEOCODE_ID	nominal	Integer	CatalogTRAININGCS_CLIE...	☑	0	0	
19	GENDER	GENDER	nominal	String	CatalogTRAININGCS_CLIE...	☑	0	0	
20	BIRTH_DATE	BIRTH_DATE	continuous	Date	CatalogTRAININGCS_CLIE...	☑	0	0	
21	BRANCHID	BRANCHID	continuous	Integer	CatalogTRAININGCS_CLIE...	☑	0	0	
22	R	R	continuous	Integer	CatalogTRAININGCS_CLIE...	☑	0	0	
24	GEOCODE_ID	GEOCODE_ID	nominal	Integer	CatalogTRAININGCS_ACC...	☑	0	0	
25	FREQUENCY	FREQUENCY	nominal	String	CatalogTRAININGCS_ACC...	☑	0	0	
26	ACCOUNT_DATE	ACCOUNT_DATE	continuous	Date	CatalogTRAININGCS_ACC...	☑	0	0	

Time-stamped Population

	Alias	Name	Type	Storage	Source	Visibility	Order	Key	Description
1	KxTimeStamp	KxTimeStamp	nominal	Date	Function	☑	0	1	
2	HAS_LOAN_EXST_ID	HAS_LOAN_EXST_ID	nominal	Boolean	Aggregate	☐	0	0	Exists(ID)
3	TARGET	TARGET	nominal	Integer	Condition	☐	0	0	
4 🔒	KxId	KxId	nominal	Integer	DISP_ID_1	☑	0	1	
5	CLIENT_ID	CLIENT_ID	continuous	Integer	DISP_ID_1	☐	0	0	
6	ACCOUNT_ID	ACCOUNT_ID	continuous	Integer	DISP_ID_1	☐	0	0	
7	CLIENT_TYPE	CLIENT_TYPE	nominal	String	DISP_ID_1	☐	0	0	
8	ID	ID	continuous	Integer	CatalogTRAININGCS_LOAN	☐	0	1	
9	ACCOUNT_ID_1	ACCOUNT_ID	continuous	Integer	CatalogTRAININGCS_LOAN	☐	0	0	
10	DURATION	DURATION	ordinal	Integer	CatalogTRAININGCS_LOAN	☐	0	0	
11	STATUS	STATUS	nominal	String	CatalogTRAININGCS_LOAN	☐	0	0	
12	AMOUNT	AMOUNT	continuous	Integer	CatalogTRAININGCS_LOAN	☐	0	0	
13	PAYMENT	PAYMENT	continuous	Integer	CatalogTRAININGCS_LOAN	☐	0	0	
14	DATE	DATE	continuous	Date	CatalogTRAININGCS_LOAN	☐	0	0	



Filter

Model Overview

Overview

Model: TARGET_CS4_TSP_LOAN2

Data Set:	BANK_AR_3_CS4_TSP_LOAN_1649147309
Analytical Record:	BANK_AR_3
Time-stamped Population:	CS4_TSP_LOAN
Entity:	DISP_ID
Reference Date:	2014-01-01
Initial Number of Variables:	46
Number of Selected Variables:	68
Number of Records:	4,135
Building Date:	2022-04-05 16:29:39
Learning Time:	15 s
Engine Name:	Kxen.RobustRegression
Author:	train-04

Nominal Targets

TARGET

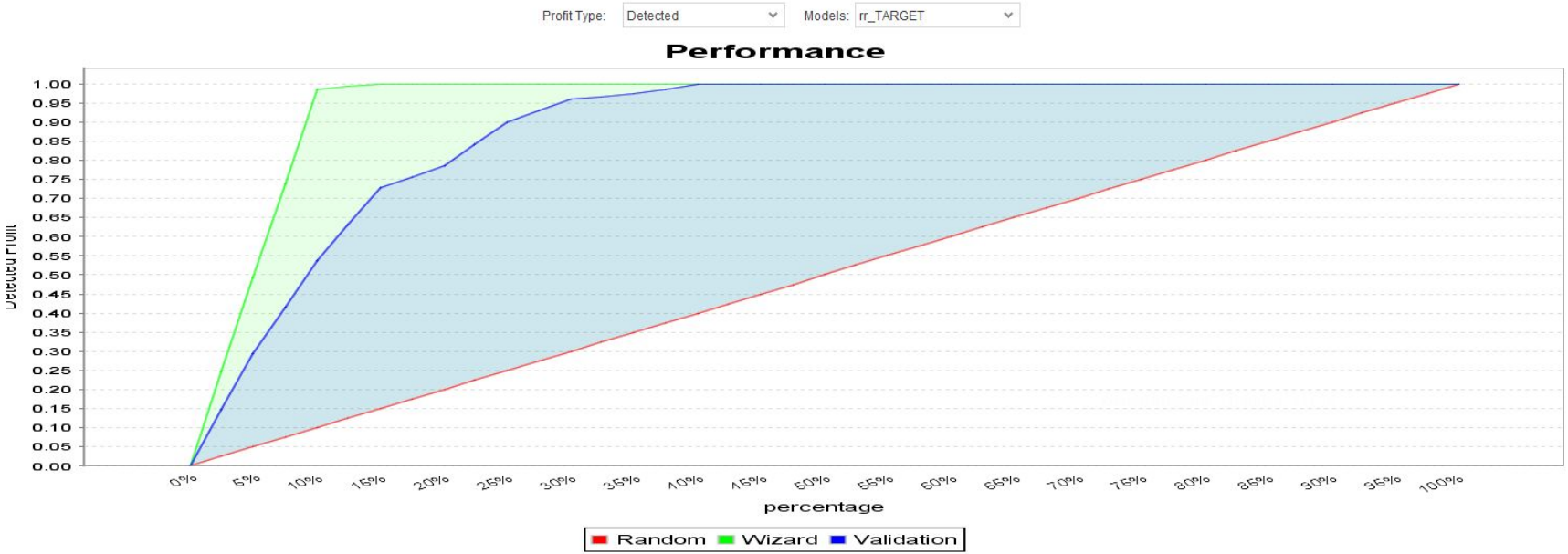
Target Key	1
0 - Frequency	91.17%
1 - Frequency	8.83%

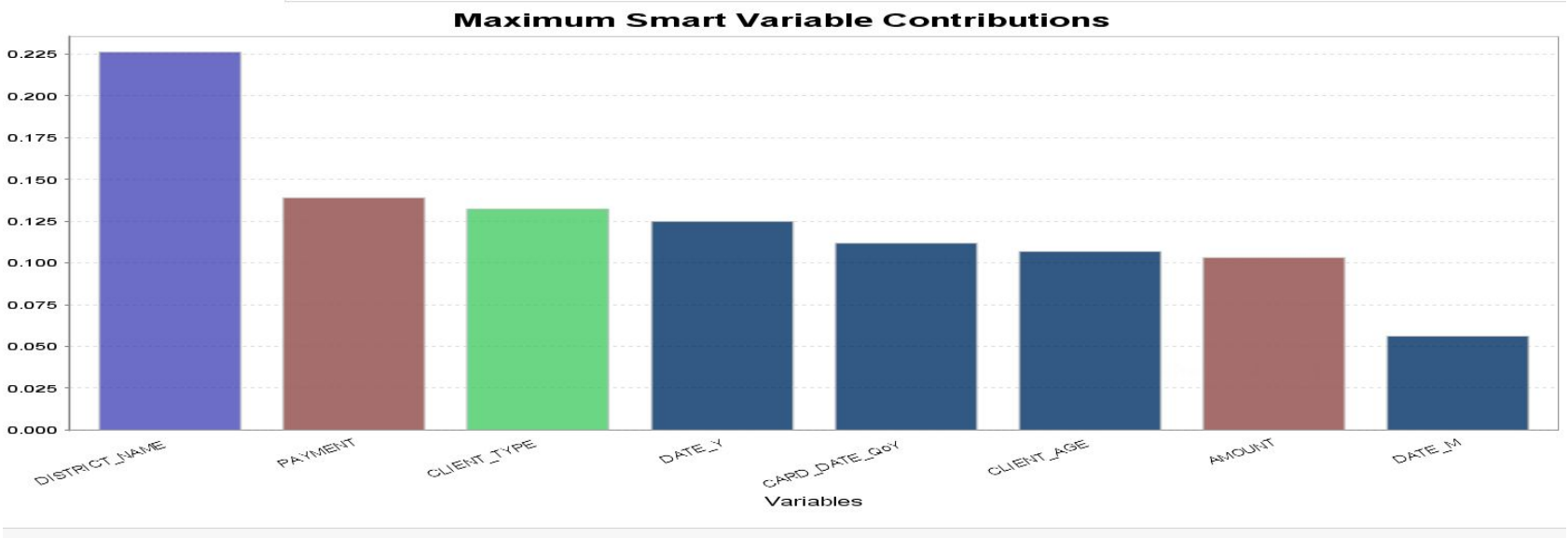
Selection Process Selected Iteration

4

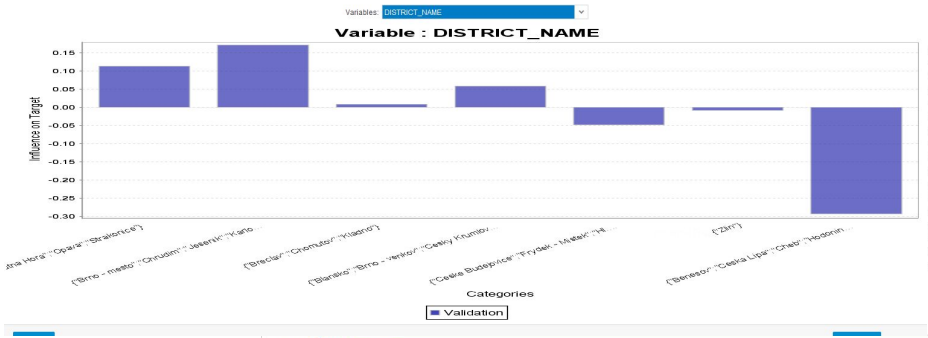
Predictive Power (KI)	0.8577
Prediction Confidence (KR)	0.9595
Nb. Variables Kept	19

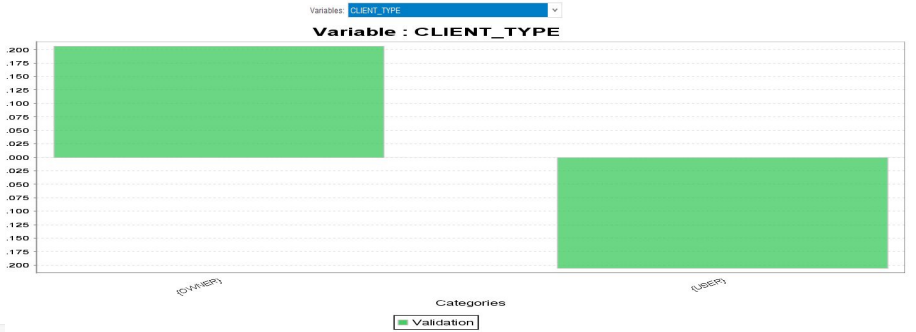
Model Performance Graph

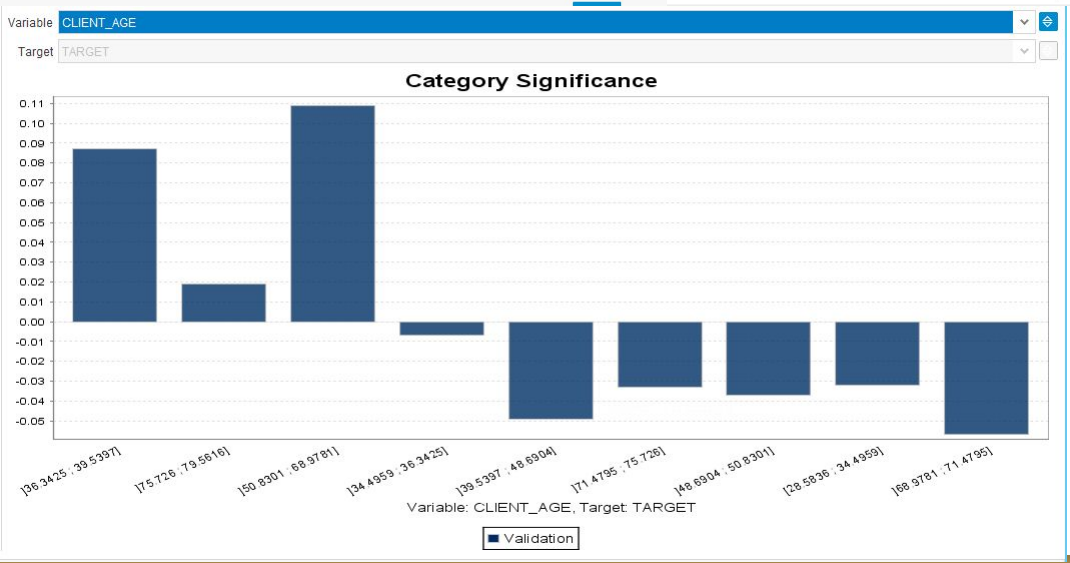



Maximum Smart Variable Contributions
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
DISTRICT_NAME
PAYMENT
CLIENT_TYPE
DATE_Y
CARD_DATE_GoY
CLIENT_AGE
AMOUNT
DATE_M
Variables

Category Significance







Model Performance Indicators

Estimator rr_TARGET

Data Set	Min	Max	Mean	Standard Deviation
Validation	-0.511	0.912	0.127	0.237
Estimation	-0.515	0.911	0.132	0.239

Indicator	Estimation	Validation
L1	0.257	0.258
L2	0.312	0.314
LInf	0.89	0.892
ErrorMean	-0.132	-0.114
ErrorStdDev	0.282	0.292
ClassificationRate	0.904	0.898

Data Set Mean and Standard Deviation between the Validation and Estimation are close.

Confusion Matrix

Confusion Matrix

	Predicted 1 (117)	Predicted 0 (1007)
True 1 (114)	63 5.6%	51 4.54%
True 0 (1010)	54 4.8%	956 85.05%

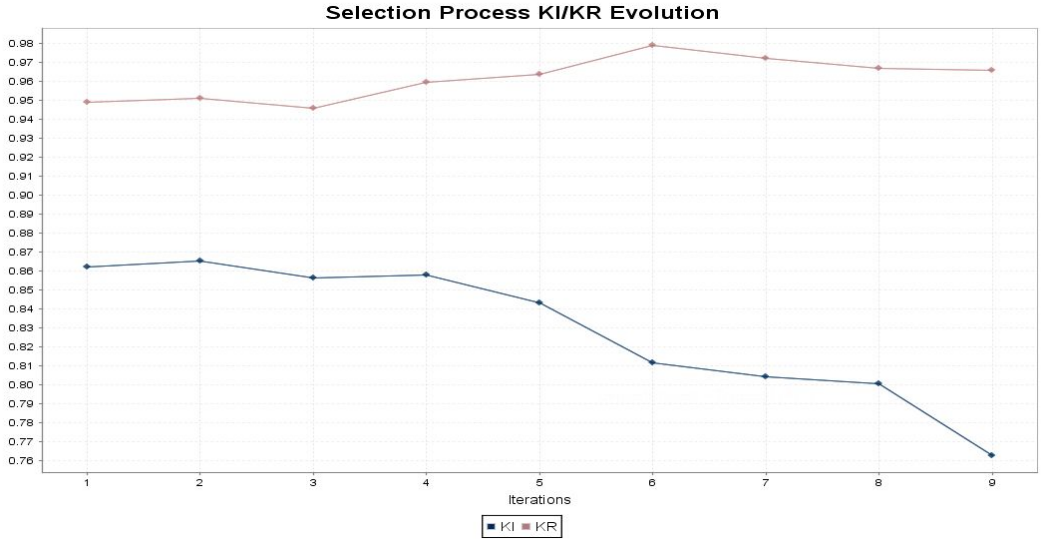
Total Population: 1,124

Actual Values

Positive(1) Negative(0)

PredictedValues

Positive(1)	TP	FP
Negative(0)	FN	TN



Iterations	KI	KR	Nb. Variables Kept
1	0.8622	0.9489	53
2	0.8655	0.9509	39
3	0.8564	0.9458	28
4	0.8577	0.9595	19
5	0.8434	0.9638	12
6	0.8117	0.9791	8
7	0.8045	0.9723	7
8	0.8007	0.9666	6
9	0.7626	0.9659	5

Analyze Deviations

Data Set used for Deviation Control	
Data Set:	FromModel_CS4_TSP_LOAN_1649668228
Source:	HDB64
Number of Records:	4,135
Number of Variables Showing Deviation:	

Deviant Variables

Performance on Control Data Set

Target: TARGET

rr_TARGET	
KI	0.8849
KR	0.9722

Apply Model KI & KR are similar to the Original model KI & KR.The model using Apply Data Set is robust and can be applied safely

Model Output in HANA

Your model was applied successfully

Summary	
Model:	TARGET_CS4_TSP_LOAN
Input:	HDB64/CS4_TSP_LOAN (Kxen.ODBCStore)
Output:	HDB64/"STUDENT04"."CS4_LOAN_DEFAULETRS4_OUTPUT" (Kxen.ODBCStore)

We have applied the model and saved the output in SAP HANA and viewed the output data

KxId	KxTimeStamp	ID	CARD_ID	TARGET	rr_TARGET
68,494	Jan 1, 2014	39,129	7,175	0	-0.114568156847...
68,635	Jan 1, 2014	39,154	7,187	0	0.138631616503 5...
68,781	Jan 1, 2014	39,167	7,194	0	0.1115513222634...
116	Jan 1, 2014	4,986	16	0	0.0738113090468...
127	Jan 1, 2014	4,989	17	0	0.3682059174699...
992	Jan 1, 2014	5,134	164	0	-0.025640732640...
1,197	Jan 1, 2014	5,154	196	0	0.0036464376793...
2,267	Jan 1, 2014	5,332	355	0	-0.341746666427...
2,606	Jan 1, 2014	5,388	414	0	0.3714683708740...
3,291	Jan 1, 2014	5,526	517	0	-0.189490119245...
3,906	Jan 1, 2014	5,634	612	0	0.0106907632098...
3,984	Jan 1, 2014	5,644	623	0	0.1473738149273...
5,224	Jan 1, 2014	5,861	769	0	0.0519376322023...
5,253	Jan 1, 2014	5,868	775	0	0.3971830691568...
5,468	Jan 1, 2014	5,906	785	0	-0.004835292174...
5,526	Jan 1, 2014	5,923	792	0	-0.173624752652...
5,978	Jan 1, 2014	6,007	819	0	-0.061971646098...
6,196	Jan 1, 2014	6,044	824	0	-0.081060794000...
6,299	Jan 1, 2014	6,063	833	1	0.4162862000195...
6,610	Jan 1, 2014	6,118	847	0	0.0881886935615...
6,753	Jan 1, 2014	6,144	854	0	-0.067946285722...
7,434	Jan 1, 2014	6,254	891	0	-0.089113465708...
7,567	Jan 1, 2014	6,278	896	0	-0.034278304590...
7,576	Jan 1, 2014	6,279	897	0	-0.217854705978...
7,672	Jan 1, 2014	6,294	904	0	0.0745075860575...
7,796	Jan 1, 2014	6,313	918	0	0.3343571425322...
7,855	Jan 1, 2014	6,321	923	0	-0.489884416719...
8,030	Jan 1, 2014	6,346	936	0	-0.087791218248...

Scenario 4: Predict Loan Defaulters

Conclusion

KI = 0.8577 ✓
KR = 0.9595 ✓

(KI > 0.45)
(KR > 0.95)

Bank reduces the number of bad loans from 11% to 7% ✓

Scenario 5:Predict Propensity to Buy Loan Products

Scenario 5: Promote Property Loan

Business Goals:

- Analyze customers to design an appropriate property loan product and communications strategies for each group. 4 to 7 groups
- Establish better customer relationship management strategies
- Increase overall customer satisfaction and loyalty by improving existing services in other departments

Business Success Criteria:

- Cross sales increase by 8%
- Customer click-through rate on promotional offers increase from 6% to 10%

Data Science Goals and Success Criteria:

- Model should have clusters usable by marketing team
- Clusters should not have too few or overly large numbers
- Predictive Power (KI) should be high (>0.58)
- Prediction Confidence (KR > 0.95)

Scenario 5: Promote Property Loan

Model Overview

- Build 4 to 7 models
- PA chose 7 clusters as optimum. KI = 0.7884, KR = 0.9951
- Note that cluster 7 has lowest unassigned record at 0.38%

Target: ANNUAL_INCOME

Model	Winner	KI	KR	Initial Number of Clusters	Final Number of Clusters	Overlap	Percentage of Unassigned Records
Engine0	false	0.7546	0.9914	4	4	12.69%	2.62%
Engine1	false	0.7849	0.9913	5	5	13.89%	1.92%
Engine2	false	0.7801	0.9922	6	6	17.22%	0.84%
Engine3	true	0.7884	0.9951	7	7	27.8%	0.38%
Engine4	false	0.7778	0.9945	8	8	29.62%	1.09%

Scenario 5: Promote Property Loan

Model Performance

Lowest frequency

Highest frequency

Cluster Name	Frequencies	Target Means
2	11.56%	7250.22
3	34.83%	2814.62
4	7.12%	8260.21
5	8.09%	6154.2
6	14.61%	21.6736
7	11.08%	11498.5

Lowest Target Mean

Highest Target Mean

Scenario 5: Promote Property Loan

Kullback-Liebler (KL) Statistic

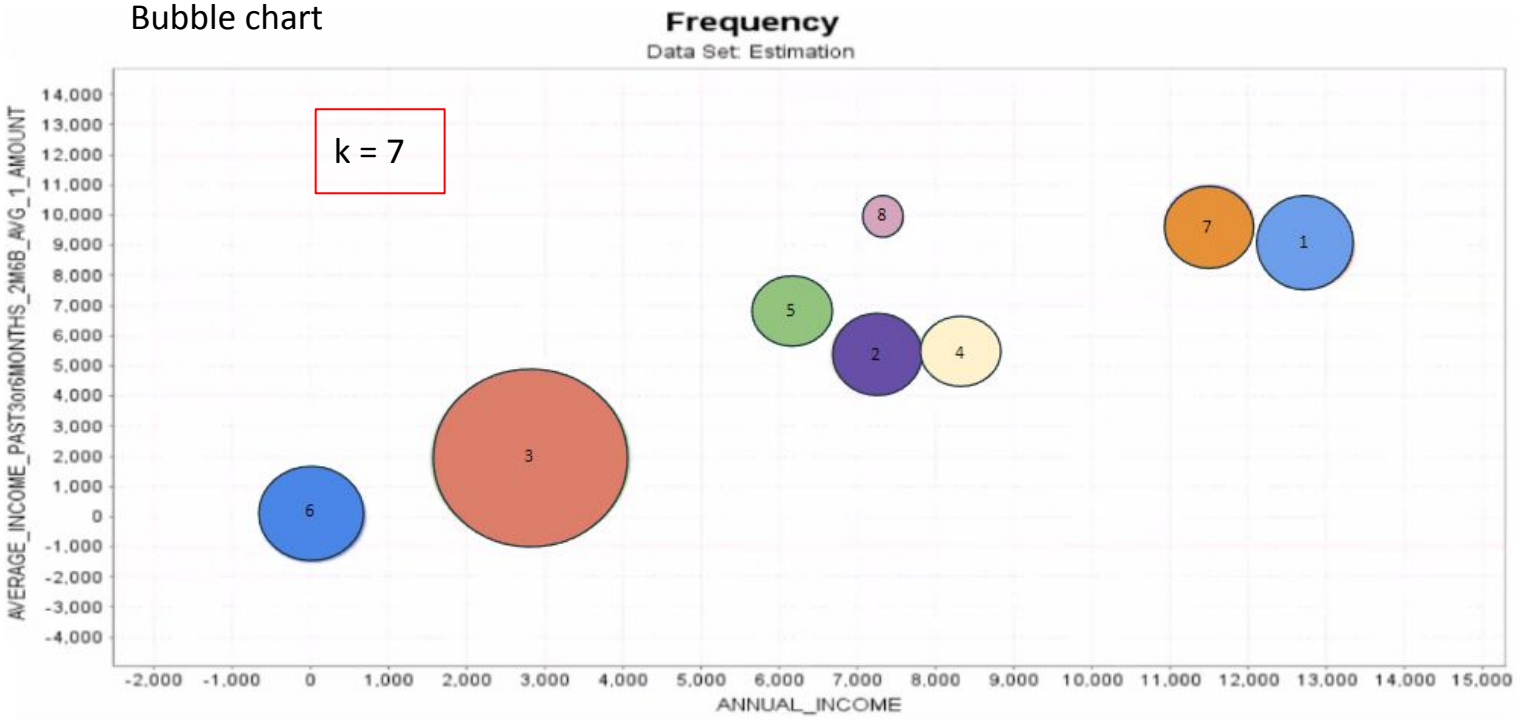
Measure divergence between variable distribution in population and in cluster.
The higher it is, the more the distributions differ.

Variable	1 - KL	2 - KL	3 - KL	4 - KL	5 - KL	6 - KL	7 - KL
AVERAGE_INCOME_...	0.888	1.372	0.639	0.532	0.860	1.637	1.580
AVERAGE_INCOME_...	0.749	0.907	0.971	0.355	0.882	1.918	0.859
MAX_BALANCE_1M6...	1.535	0.756	0.790	1.163	0.301	1.924	0.845
AVERAGE_BALANCE...	0.625	0.278	0.277	0.411	0.315	1.697	0.396
MIN_ACCOUNT_BAL...	0.625	0.278	0.277	0.411	0.315	1.697	0.396
AVERAGE_BALANCE...	0.697	0.300	0.288	0.455	0.414	1.630	0.390
MIN_ACCOUNT_BAL...	0.697	0.300	0.288	0.455	0.414	1.630	0.390
AVERAGE_BALANCE...	0.664	0.276	0.319	0.445	0.272	1.807	0.335
MIN_ACCOUNT_BAL...	0.664	0.276	0.319	0.445	0.272	1.807	0.335
AVERAGE_BALANCE...	0.711	0.234	0.390	0.439	0.226	1.909	0.296
AVERAGE_BALANCE...	0.902	0.280	0.315	0.575	0.449	1.528	0.285
MIN_ACCOUNT_BAL...	0.902	0.280	0.315	0.575	0.449	1.528	0.285
MIN_BALANCE_1M6...	0.661	0.192	0.299	0.492	0.719	1.924	0.259
AVERAGE_BALANCE...	0.975	0.308	0.244	1.126	0.153	1.065	0.188

Value > 1 suggests KL measure is significant and reliable

Scenario 5: Promote Property Loan

Bubble chart



Scenario 5: Promote Property Loan

4 clusters
KI = 0.6752
KR = 0.9849

Continuous Targets (Number)

ANNUAL_INCOME	
Min	0
Max	27,081.3
Mean	5,833.89
Standard Deviation	5,699.34

Cluster Counts

ANNUAL_INCOME	
Initial Number of Clusters	4
Final Number of Clusters	4

Target: ANNUAL_INCOME

kc_ANNUAL_INCOME	
Predictive Power (KI)	0.6752
Prediction Confidence (KR)	0.9849

Scenario 5: Promote Property Loan

Cluster frequency, Target Mean and Target Standard Deviation

Highest frequency

Lowest frequency

Cluster	Frequency	Target Mean	Target Standard Deviation
1	23.83%	9,571.37	5,685.69
2	16.11%	299.094	1,365.8
3	42.07%	3,704.24	2,513.7
4	17.99%	10,817.2	6,316.91

Highest Target Mean

Lowest Target Mean

Scenario 5: Promote Property Loan

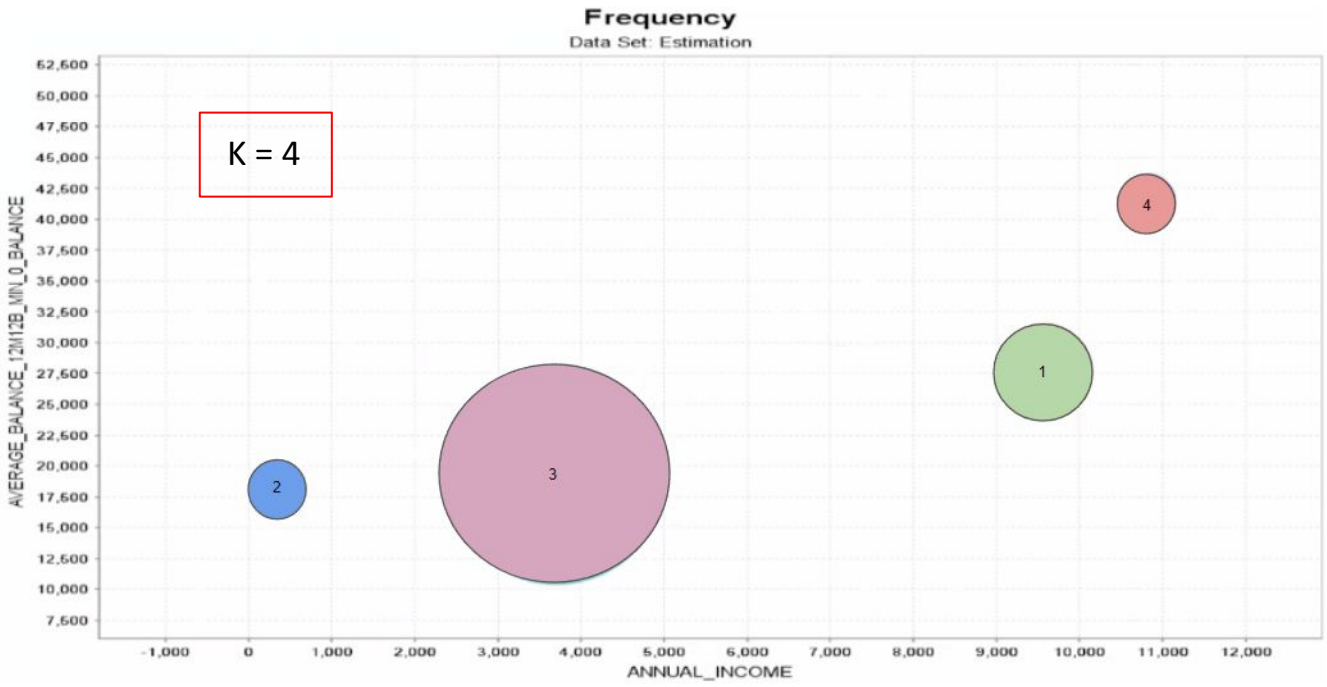
Kullback-Liebler (KL) Statistic

Variable	1 - KL	2 - KL	3 - KL	4 - KL
MAX_BALANCE_1M6B...	1.576	0.842	1.345	1.924
AVERAGE_INCOME_P...	0.487	0.521	0.711	1.918
AVERAGE_BALANCE_...	0.322	0.426	0.533	1.909
MIN_ACCOUNT_BALA...	0.322	0.426	0.533	1.909
AVERAGE_BALANCE_...	0.345	0.353	0.520	1.807
MIN_ACCOUNT_BALA...	0.345	0.353	0.520	1.807
AVERAGE_BALANCE_...	0.313	0.277	0.491	1.697
MIN_ACCOUNT_BALA...	0.313	0.277	0.491	1.697
AVERAGE_INCOME_P...	0.497	0.560	0.727	1.637
AVERAGE_BALANCE_...	0.316	0.274	0.484	1.630
MIN_ACCOUNT_BALA...	0.316	0.274	0.484	1.630
AVERAGE_BALANCE_...	0.297	0.286	0.537	1.528
MIN_ACCOUNT_BALA...	0.297	0.286	0.537	1.528
AVERAGE_BALANCE_...	0.230	0.277	0.491	1.413
MIN_ACCOUNT_BALA...	0.230	0.277	0.491	1.413
AVERAGE_BALANCE_...	0.218	0.261	0.422	1.324
MIN_ACCOUNT_BALA...	0.218	0.261	0.422	1.324
AVERAGE_BALANCE_...	0.195	0.247	0.374	1.213
MIN_ACCOUNT_BALA...	0.195	0.247	0.374	1.213
AVERAGE_BALANCE_...	0.187	0.248	0.339	1.138
MIN_ACCOUNT_BALA...	0.187	0.248	0.339	1.138
AVERAGE_BALANCE_...	0.240	0.216	0.273	1.065
MIN_ACCOUNT_BALA...	0.240	0.216	0.273	1.065
AVERAGE_BALANCE_...	0.147	0.201	0.278	0.978
MIN_ACCOUNT_BALA...	0.147	0.201	0.278	0.978

Value > 1 suggests KL measure is significant and reliable

Scenario 5: Promote Property Loan

Bubble chart



Scenario 5: Promote Property Loan

Strategies and grouping

Cluster 1 (Affluent Customers)

- To initiate loans, provide attractive mortgage interest rate
- Take a moderate credit risk approach
- Build relationships

Cluster 2 (Younger Customers)

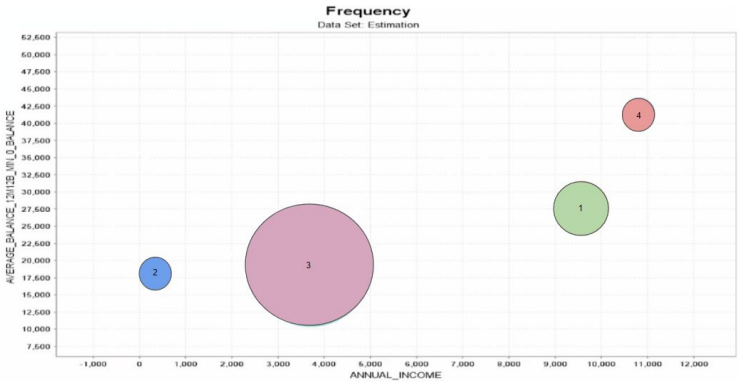
- Use digital tools, integrate automation
- Customer outreach programme

Cluster 3 (Mass Customers)

- Take a tight but controlled credit risk approach
- Make use of social media strategies

Cluster 4 (Wealthy Customers)

- Take a high credit risk approach
- Offer white glove service



Scenario 5: Promote Property Loan

Conclusion

KI = 0.6752 ✓
KR = 0.9849 ✓

(KI > 0.58)
(KR > 0.95)

4 well distributed clusters chosen and usable by marketing team. ✓
Reasonable sized

Scenario 6: Forecast Deposit Balances

Scenario 6: Predict Deposits

Business Goals:

- Use Predictive Analytics tools to estimate the amount the customers will deposit for the next three months

Business Success Criteria:

- Deposit amount increases by 5% over last year's deposits
- Bank can service 75% of its loan requests

Data Science Goals and Success Criteria:

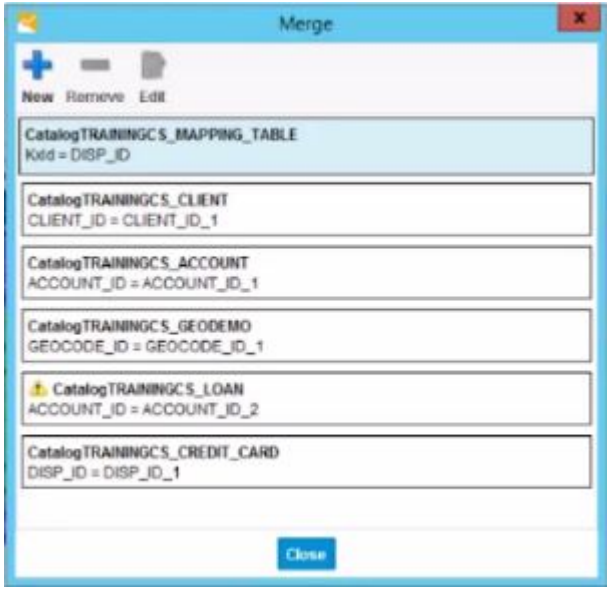
- Create a regression model that estimates the deposit sum for each customer for the next three months using one month latency, six months history and three months target period
- Predictive Power (KI) should be high (>0.65)
- Prediction Confidence (KR > 0.95)

Scenario 6: Predict Deposits

Regression Model

1

Data Preparation - Merging



2

Enhancing Analytical Record

7	AVERAGE_BALANCE_4M12B...	AVERAGE_BALANCE_4M12B...	continuous
8	AVERAGE_BALANCE_4M12B...	AVERAGE_BALANCE_4M12B...	continuous
9	AVERAGE_PAYMENT_4M12B...	AVERAGE_PAYMENT_4M12B...	continuous
10	AVERAGE_PAYMENT_4M12B...	AVERAGE_PAYMENT_4M12B...	continuous
11	AVERAGE_PAYMENT_4M12B...	AVERAGE_PAYMENT_4M12B...	continuous
12	AVERAGE_PAYMENT_4M12B...	AVERAGE_PAYMENT_4M12B...	continuous
13	MIN_PAYMENT_4M12B_MIN_...	MIN_PAYMENT_4M12B_MIN_...	continuous
14	MIN_PAYMENT_4M12B_MIN_...	MIN_PAYMENT_4M12B_MIN_...	continuous
15	MIN_PAYMENT_4M12B_MIN_...	MIN_PAYMENT_4M12B_MIN_...	continuous
16	MIN_PAYMENT_4M12B_MIN_...	MIN_PAYMENT_4M12B_MIN_...	continuous
17	MAX_PAYMENT_4M12B_MAX...	MAX_PAYMENT_4M12B_MAX...	continuous
18	MAX_PAYMENT_4M12B_MAX...	MAX_PAYMENT_4M12B_MAX...	continuous
19	MAX_PAYMENT_4M12B_MAX...	MAX_PAYMENT_4M12B_MAX...	continuous

Variables

- Minimum Balance, Maximum Balance, Average income over past 3 months, month-on-month income ratio
- Population must be with bank longer than 6 months

Scenario 6: Predict Deposits

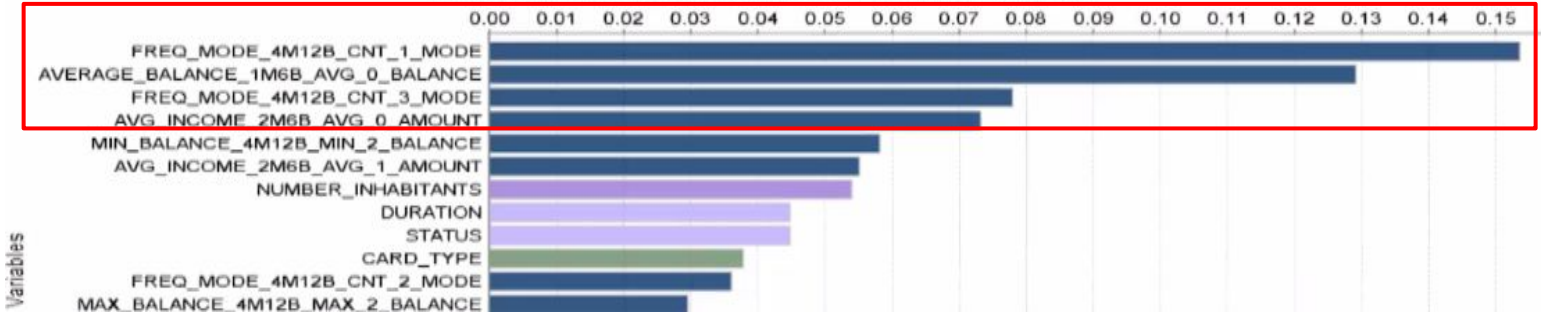
Continuous Targets (Number)

TARGET_DEPOSIT_NEXT_3_MONTHS	
Min	658.8
Max	220,932
Mean	83,968.7
Standard Deviation	44,860

Selection Process Selected Iteration

2	
Predictive Power (KI)	0.9612
Prediction Confidence (KR)	0.9717
Nb. Variables Kept	22

Maximum Smart Variable Contributions



Scenario 6: Predict Deposits

Issues

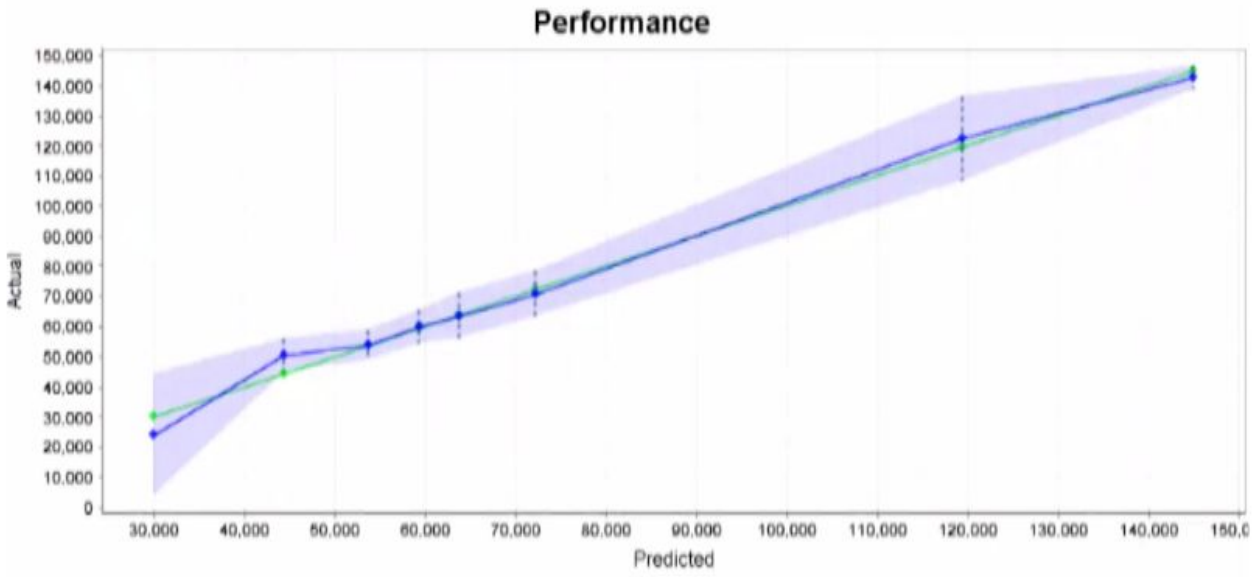
Variable	Category	Validation - Probability of Deviation
AVG_INCOME_2M6B_AVG_1_AMOUNT	]9933.63 ; 10976.6]	0.965
AVERAGE_SALARY	]8843 ; 8994]	0.981
TARGET_DEPOSIT_NEXT_3_MONTHS	]50412.5 ; 59503.5]	0.993
MAX_BALANCE_4M12B_MAX_2_BALANCE	]49442.4 ; 56329.1]	0.984
MIN_BALANCE_4M12B_MIN_0_BALANCE	]23292.7 ; 26389.9]	0.988
NUMBER_INHABITANTS	]197099 ; 228848]	0.981
MIN_BALANCE_4M12B_MIN_2_BALANCE	[6998.6 ; 20302.9]	0.984
NUMBER_CITIES	[4]	0.953
FREQ_MODE_4M12B_CNT_0_MODE	[14]	0.984
RATIO_ENTREPRENEURS	[110]	0.995

Indicators

Indicator	Estimation	Validation
L1	15,019.7	14,954.5
L2	30,077.7	30,507.7
LInf	320,005	237,353
ErrorMean	93.559	108.23
ErrorStdDev	30,077.6	30,507.5
R2	0.566	0.51

Scenario 6: Predict Deposits

Model Performance



Purple area around the graph shows where actual values are expected to be. This is the confidence interval

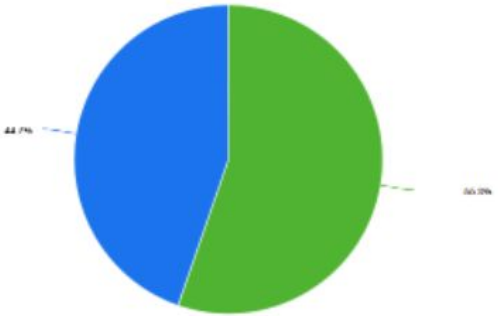
Scenario 6: Predict Deposits

BI Reporting

Analysis | Distinct values | Raw Data

Filter pattern | 55 rows retrieved - 28 ms

MOBILE_NUMBERS	CLIENT_ID	GENDER	REGION	AGE	CITY	EMAIL	BRANCHID
888600964	964	F	New York	63	New York	dummyem9...	10
888202632	2,632	M	California	43	San Jose	dummyem2...	14
888503678	3,678	F	California	48	Los Angeles	dummyem3...	1
888809539	9,539	M	Texas	38	Austin	dummyem1...	12
888210059	10,059	M	New York	38	New York	dummyem9...	10
888210059	10,059	F	California	37	Los Angeles	dummyem1...	41



Scenario 6: Predict Deposits

Results

KI = 0.9612 ✓
KR = 0.9717 ✓

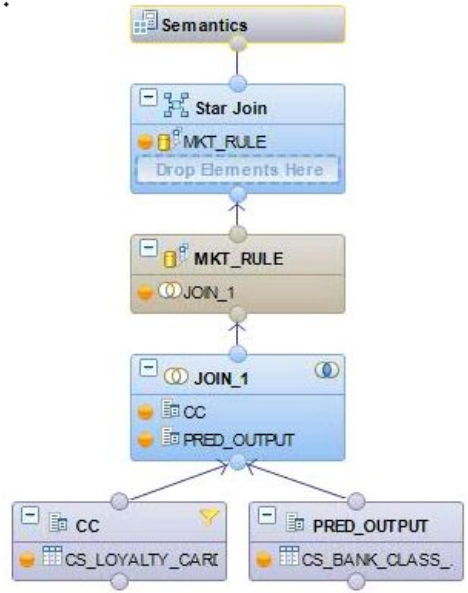
(KI > 0.65)
(KR > 0.95)

Deposit amount increased by 5% ✓

Bank can service 75% of loans ✓

Understand the Background

VW_PREDICT_BUY



	JOIN_1	MKT_RULE	STAR JOIN
Tables	CS_BANK_CLASS_OUTPUT CS_LOYALTY_CARD	JOIN_1	MKT_RULE
Joins	INNER		
Cardinality	1:1		
Filter	“AGE_GROUP”!=’-’(CC NODE)		
Calculated Columns	KXID_INT(PRED_OUTPUT NODE)	CLEAN_AGE_GROUP AVG_PROBA_DECISION AVG_PROBA_DECISION_TGT MIN_AGE MAX_AGE	
Output	ACTUAL & PREDICT TARGET	ACTUAL & PREDICT TARGET	ACTUAL & PREDICT TARGET

Calculated Column	KXID_INT	Int(“KxId”)	To convert KxId to integer data type
Calculated Column	CLEAN_AGE_GROUP	CASE WHEN”AGE_GROUP”=’?? 17 THEN ‘0-17’ WHEN “AGE_GROUP”=’????? 60’ THEN 60-200 ELSE “AGE_GROUP” END	To rename age groups
Calculated Column	AVG_PROBA_DECISION	“proba_decision_rr_TARGET”/”RECCNT”	To calculate the average probability for each age group
Calculated Column	AVG_PROBA_DECISION_TGT	“decision_rr_TRAGET”/”RECCNT”	To calculate the average target value for each age group
Calculated Column	MIN_AGE	leftsrtr(“CLEAN_AGE_GROUP”,2)	To get minimum age of the age group
Calculated Column	MAX_AGE	rightstr(CLEAN_AGE_GROUP”,3)	To get the maximum age of the age group

Raw Data | Distinct values | Analysis

Filter pattern

200 rows retrieved - 32 ms

KxId	KxTimeStamp	TARGET	rr_TARGET	decision_rr_TARGET	proba_decision_rr_TARGET
2,193,397	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
2,267,593	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
2,530,109	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
71,925	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
256,469	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
369,145	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
2,387,093	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
2,793,085	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
3,052,037	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
339,049	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
479,301	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
1,854,305	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
1,851,489	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
18,661	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
92,005	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228
307,765	Jul 1, 2014	0	0.2232048654997...	0	0.9952989237395228

CS_BANK_CLASS_OUTPUT

Analysis | Distinct values | Raw Data

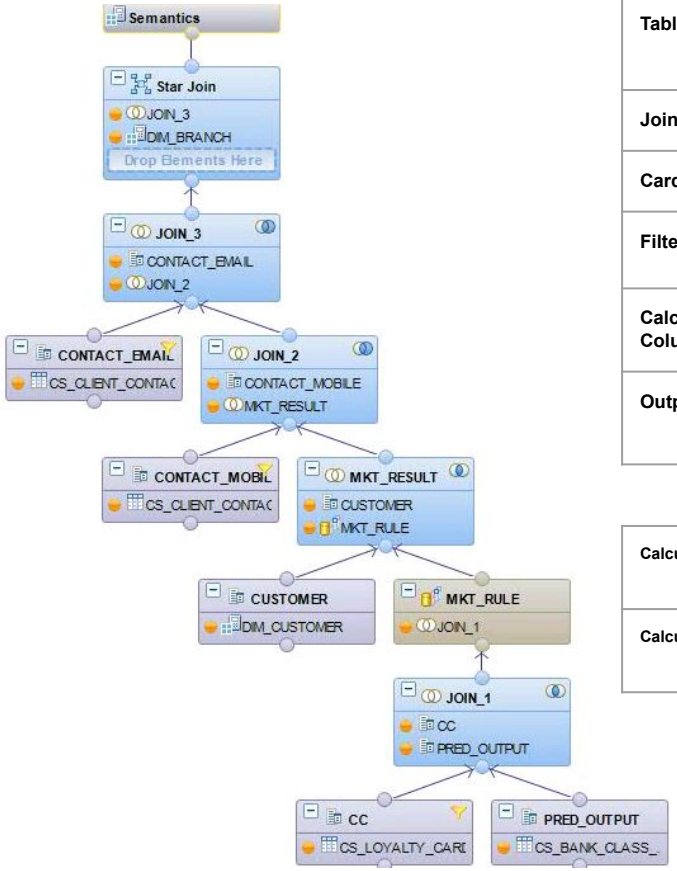
Filter pattern

12 rows retrieved - 109 ms

GENDER_TYPE	CLEAN_AGE_GROUP	MIN_AGE	MAX_AGE	RECCNT	AVG_PROBA_DECISION	AVG_PROBA_DECISION_TGT
F	18 - 25	18	25	5,537	0.983581604390481	0.005598
M	18 - 25	18	25	3,642	0.9842897388827977	0.005216
F	26 - 35	26	35	46,718	0.9802418552917337	0.0076619999999999995
M	26 - 35	26	35	31,076	0.9803102411490062	0.007658
F	36 - 45	36	45	47,268	0.9799268387257933	0.007424999999999999
M	36 - 45	36	45	31,839	0.979544219588136	0.007883
F	46 - 60	46	60	35,555	0.9813434889832106	0.006018
M	46 - 60	46	60	23,877	0.9812883697643754	0.00603
F	0-17	0	-17	18,857	0.982458587078798	0.006469
M	0-17	0	-17	12,614	0.9826148388641561	0.006104
F	60-200	60	200	12,062	0.9838979414060883	0.0038959999999999997
M	60-200	60	200	8,048	0.9846819412531985	0.002982

VW_PREDICT_BUY

VW_PREDICT_BUY_CAMPAIGN



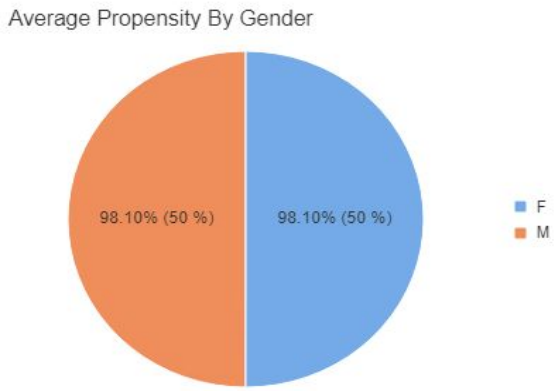
	MKT_RESULTS	JOIN_2	JOIN_3	STAR JOIN
Tables	MKT_RULE DIM_CUSTOMER	MKT_RESULTS CS_CLIENT_CONTAC T	JOIN_2 CS_CLIENT_CONTACT	JOIN_3 DIN_BRANCH
Joins	Inner	Left Outer	Left Outer	Inner
Cardinality	1:1	1:1	1:1	1:1
Filter		“CONTACTTYPE”=MO BILE	“CONTACTTYPE”=EMAIL	
Calculated Column	AGE DUM			
Output	ACTUAL & PREDICT TARGET	ACTUAL & PREDICT TARGET	ACTUAL & PREDICT TARGET	ACTUAL & PREDICT TARGET

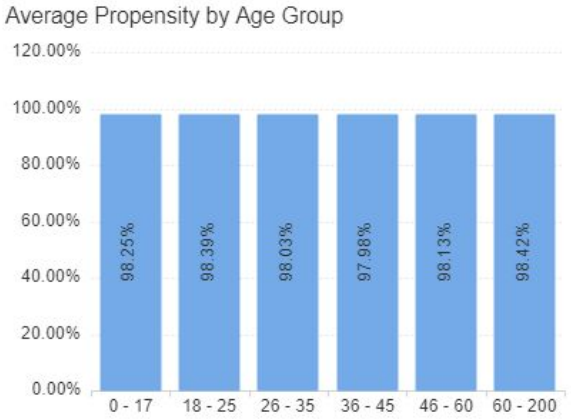
Calculated Column	AGE	(days_between(“BIRTH_DATE”,current_date))/365.25	To calculate the age of card member
Calculated Column	DUM	1	Dummy field for join

Data Preview

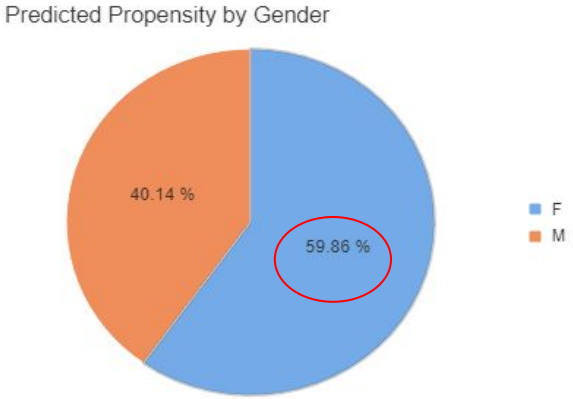
AGE	CLEAN_AGE_GROUP	MIN_AGE	MAX_AGE	GENDER	MOBILE	EMAIL	CLIENT_ID	RECCNT	AVG_PROBA_DECISION	AVG_PROBA_DECISION_TGT
66	46 - 60	46	60	F	888859663	dummyem5...	59,663	35,555	0.9813434889831995	0.006018
83	0-17	0	-17	F	888143635	dummyem4...	43,635	18,857	0.9824585870788037	0.006469
91	36 - 45	36	45	M	888116107	dummyem1...	16,107	31,839	0.9795442195881392	0.007883
68	60-200	60	200	F	888031038	dummyem3...	31,038	12,062	0.9838979414060882	0.0038959999999999997
43	0-17	0	-17	M	888552636	dummyem5...	52,636	12,614	0.9826148388641527	0.006104
63	46 - 60	46	60	M	888217913	dummyem1...	17,913	23,877	0.9812883697643735	0.00603
95	36 - 45	36	45	M	888502455	dummyem2...	2,455	31,839	0.9795442195881392	0.007883
64	46 - 60	46	60	F	888343668	dummyem4...	43,668	35,555	0.9813434889831995	0.006018
69	26 - 35	26	35	F	888643793	dummyem4...	43,793	46,718	0.9802418552917351	0.0076619999999999995
62	18 - 25	18	25	M	888006191	dummyem6...	6,191	3,642	0.9842897388827979	0.005216
84	60-200	60	200	M	888614355	dummyem1...	14,355	8,048	0.9846819412531979	0.002982
69	60-200	60	200	M	888029714	dummyem2...	29,714	8,048	0.9846819412531979	0.002982
66	46 - 60	46	60	F	888845881	dummyem4...	45,881	35,555	0.9813434889831995	0.006018
62	36 - 45	36	45	M	888258845	dummyem5...	58,845	31,839	0.9795442195881392	0.007883
84	26 - 35	26	35	F	888159580	dummyem5...	59,580	46,718	0.9802418552917351	0.0076619999999999995
56	26 - 35	26	35	F	888403921	dummyem3...	3,921	46,718	0.9802418552917351	0.0076619999999999995
58	36 - 45	36	45	F	888328228	dummyem2...	28,228	47,268	0.9799268387257879	0.007424999999999999
66	18 - 25	18	25	M	888815761	dummyem1...	15,761	3,642	0.9842897388827979	0.005216
56	46 - 60	46	60	F	888321435	dummyem2...	21,435	35,555	0.9813434889831995	0.006018

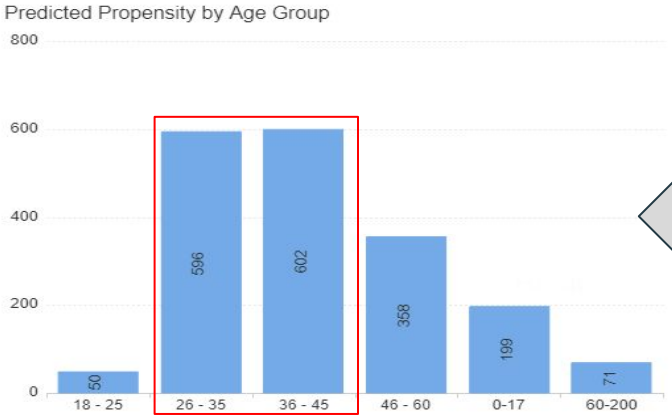
Propensity Analysis





Average Predicted Propensity by Age & Gender





Predicted Propensity by Age & Gender

Overall Conclusion - Business Objectives

Business goal and objectives

- Reduce risk from bank loan
- Identify defaulters at early stage
- Cross sell credit card and property loan

Classification model that helps to increase profit and reduce risk by identifying the right customers

Classification, regression and cluster model.

- Convert > 10% loyalty scheme customers to new credit card
- Inform the bank on how customers use their credit card
- Design communication strategies and property loan product to target group

Overall Conclusion - Business Objectives

Business goal and objectives

- Enhancing communications strategy
- Target marketing campaign
- Launch right products accordingly

Cluster model that helps identify groups to specifically market and design relevant strategies

Classification model and regression model to convert specific customers to new credit cards and to inform the bank on how customers use their credit card

Overall Conclusion - Business Objectives

Business goals and objectives

- Improving customer satisfaction
- Increase customer loyalty

Regression model to show how deposit amount increases by 5% over last year's figures and for the bank to serve 75% of loan requests

Thank you

FAQ